Remarks

Regarding amendments in the Claims:

Claims 6-9 and 15-19 were allowed in the Final Office Action mailed May 13, 2005. Claims 10-14 and 20-24 were rejected in that Final Office Action. In response to the rejection, applicants have amended claims 10, 20, and 24. And clarification regarding claims 14 and 24 is respectfully submitted.

Applicants have previously submitted Remarks (or Arguments) for patentability of presently pending claims in an Amendment/Response filed December 23, 2004 and a Supplemental Amendment/Response January 26, 2005. And applicants respectfully direct the Examiner to these Remarks (including the "Stamp Pasting Analogy" and Illustration on pages 18 and 19 of the Supplemental Amendment) and will, in general, not repeat those Remarks here. As is well known, support for claims need not be verbatim ("ipsis verbis" or "in haec verba"), but only described in sufficient detail that one skilled in the art can reasonably conclude that the inventor had possession of the claimed invention (see, e.g., *Vas-Cath, Inc. v. Mahurkar*, 935 F.2d at 1563, 19 USPQ2d at 1116).

Regarding claims 10 and 20 (and their dependent claims) The Examiner has rejected these claims under 35 U.S.C. 112, second paragraph as being indefinite because of the language "nearly identical" and "described in (1): (1) any one CL-F point..".
Applicants have amended claims 10 and 20 and respectfully submit that the scope of these newly amended claims is not decreased. The limitation containing the language "nearly identical" accomplishes the elimination of pairs of redundant markers in the same subset (that provide the same information). Applicants have responded and amended claims 10 and 20 by deleting the limitation containing the language "nearly identical". See [0321], which states that Step 3, the elimination of pairs of redundant markers in the same subset (that provide the same information), is not essential. Therefore, the deletion of the limitation is supported.

The language "described in (1): (1) Any one CL-F point.." has also been deleted from claims. Applicants have simplified and clarified the language in claims 10 and 20.

The CL-F region in claims 10 and 20 is now specified to be a segment-subrange, wherein the segment of the segment-subrange is the region of interest or the chromosome. (See [0275] bottom of page 20 which states that the length of a chromosomal segment can be as long as a chromosome.) Support for the limitations or language in claims 10 and 20 has been given in the Supplemental Amendment of January 26, 2005. Specifically p. 13 of that Supplemental Amendment states that any CL-F region (any collection of one or more points [0050]) is an example of a CL-F region that is systematically covered by versions of the invention. A segment-subrange is an example of such a CL-F region, see [0090]; and see [0185] ***"Specific types of CL-F regions that are N covered are useful. For example, a rectangular CL-F region, a segment-subrange,..."***.

And see top of page p. 16 of the Supplemental Amendment of Jan 2005 which gives support for the limitation *"wherein each point in the CL-F region is N-covered to within [L, y] by markers belonging to a subset, L is the length of the longest segment, y is 0.15 and N ≥ 2,"*. This limitation follows directly from the facts that the markers in each subset belong to only one segment (whose maximum length is L), the fact that the difference between the least common allele frequencies of any two subset markers does not exceed 0.15, and that there are two or more markers in each subset. See also the "Stamp Pasting Analogy" and Illustration on p. 16, 18 and 19 of the Supplemental Amendment of Jan. 2005.

Regarding claims 14 and 24 The Examiner has rejected these claims as indefinite under 35 U.S.C. 112, second paragraph because of the language "thousands of bi-allelic covering markers". Applicants respectfully offer the following clarification. The concept of "thousands of bi-allelic [covering] markers" (in connection with the physical implementation the new, two-dimensional linkage study techniques of this application) using silicon chips or glass slides containing oligonucleotides is described in [0322], [0323], and [0324]. Included in this description is the paper cited in endnote 8, that is incorporated by reference into the application (Accessing Genetic Information with High-Density DNA Arrays, Mark Chee, et al. Science, vol 274, Oct. 25, 1996, pp. 610 – 614). Other similar papers such as Large Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome, Wang, et. al., Science, May 15, 1998, vol 280, pp. 1077-1081 in endnote 9, are also incorporated by reference into the application.

Applicants respectfully submit that given this description and the knowledge of one of ordinary skill, the limitation "thousands of bi-allelic covering markers" is definite. "Thousands" is the plural of "thousand" meaning literally 2000 or more. Thousands is also a very large number. (See enclosed copy of page 1228 of the Tenth Edition of Merriam-Webster's Collegiate Dictionary giving the definition of "thousands" as plural of "thousand".) The abstract of the Chee paper (endnote 8) describes "DNA arrays containing up to 135, 000 probes" and page 610 of this paper (bottom left hand column) describes an "array of a large number of oligonucleotide probes". And the next to last paragraph of the Wang paper (endnote 9) on page 1081 describes a "2000-SNP genotyping chip".

The disadvantages of the conventional, generally slower, nucleic acid sequencing technologies (compared to high-density DNA arrays that use large numbers of oligonucleotide probes) is described in the first column, page 610 of the Chee paper. Both the Chee and Wang papers describe querying the entire human genome (estimated in Chee at 100, 000 genes) using a high-density array (p. 613 Chee and last two paragraphs p. 1081 Wang). No definite upper limit to the number of probes (and by implication number of markers) for the technology is given. It is believed that *"For example, the entire set of ~ 10^{12} 20-nucleotide oligomer probes, or any desired subset, can be synthesized...The number of probes that can be synthesized is limited only by the physical size of the array and the achievable lithographic resolution."* (first paragraph, right hand column p. 610 Chee).

The expressions "thousands of genes", "thousands of oligonucleotides", "thousands of bi-allelic markers" or similar expressions were used in the art at the time of filing of the application and are still being used. These expressions are often used in connection with high-density DNA arrays and specific example numbers (that are in the thousands). The applicants will cite several examples of this usage in the art below.

Given the extensive usage of these expressions and knowledge in the art, applicants respectfully submit the limitation "thousands of bi-allelic covering markers" is definite. In In re Corr (146 USPQ 69), the Court found that the phrase "high styrene resin" was definite and rejected the argument the phrase represented "undue breadth or overclaiming". The Court noted that the specification stated that the "high styrene resin" was a resin such as PLIOLITE S-6B. And the Court stated: *"Appellant's specification taken with the prior art clearly indicates that the styrene resin component of his composition is conventional and many equivalents are known in the art"* (146 USPQ at 71). Applicants respectfully submit that in the present application (as in In re Corr) examples of "thousands" have been given (i.e. 135, 000 probes, 100, 000 genes in the Chee paper; 2000-SNP genotyping chip in the Wang paper). And numerous equivalents of these examples of "thousands" were known in the art at the time of filing. Applicants will cite evidence that there were such numerous equivalents of "thousands" known in the art in the following two paragraphs.

Applicants respectfully direct the Examiner's attention to last paragraph on p. 772 the Fodor paper (Science, 1991, vol. 251, pp. 767-773). This paragraph describes a high-density array with 65, 536 oligonucleotides. The Fodor paper is cited as a reference in the Chee paper (note 5, p. 613). The Cann paper (C R Acad Sci III June 1998; 321(6):443-6) uses the phrase *"thousands of DNA polymorphisms (genetic markers)"* and *"thousands of more stable single nucleotide polymorphisms that detect variation on average once every ~ 1000 base pairs"* (see Abstract p. 443 and p. 445 left column bottom paragraph). The DeRisi paper (Science vol 278 October 1997 pp. 680-686) uses the phrase *"DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass slide"* (p. 680 2nd paragraph left most column). The DeRisi paper also describes the amplification of 6000 genes and microarrays with 6400 elements in each array (see p. 685 last paragraph). The Lashkari paper (Proc Natl Acad Sci USA vol 94, pp. 13057-13062 Nov. 1997) describes high density DNA arrays containing 2,479 yeast ORFs and 6, 100 ORFs (see Abstract p.13057 and next to last paragraph p. 13062). The Johnston paper (Current Biology Feb 26, 1998, 8(5) pp. R171-R174) uses the phrases "thousands of genes" and "thousands of DNA fragments" in connection with high-density DNA arrays. And Johnston also describes *"current oligonucleotide chips display all 6000 yeast genes on four 1.28 x 1.28 cm chips.. or 1.8 x 1.8 cm glass slide"* (see Abstract and second paragraph p. R171).

A Nature Genetics Supplement vol 21 January 1999 has a large number of papers on DNA microarrays. For example, the Brown paper describes *"arrays of thousands of discrete DNA sequences (for example, all 6200 known and predicted genes of S. cerevisiae"* (see p. 33 last paragraph). And the Lipshutz paper describes hundreds of thousands of oligonucleotides in an array and gives a specific number example of approximately 300,000 (see Abstract and second paragraph p. 20).

Copies of cited pages in the above papers (rather than all the pages) are included for the Examiner's convenience. These are: Chee pp. 610 and 613, Wang pp. 1077 and 1081, Fodor pp. 767 and 772, Cann pp. 443 and 445, DeRisi pp. 680 and 685, Lashkari pp. 13057 and 13062, Johnston p. R171, Brown p. 33 and Lipshutz p. 20.

As stated above, the phrase "thousands of bi-allelic markers" is included under Physical Implementation [0322] of the new, two-dimensional linkage study techniques of this application. Thousands of bi-allelic markers are thus described as a tool or implement for use by two-dimensional linkage study techniques. Indeed at the time the application was filed, the whole field of association studies is looking to use thousands of bi-allelic markers. See for example Risch, N. and Merikangas, K.: The Future of Genetic Studies of Complex Human Diseases. Science, 13 September 1996, vol. 273, pp. 1516-1517 cited in [0027] of the application. This Risch paper (see p. 1517 mid left most column) describes using technological advances to do association testing of five diallelic (or bi-allelic) polymorphisms within each of 100, 000 genes (a total of 500, 000 polymorphisms tested in the association study). And the inventor's paper is a generalization of the Risch and Merikangas analysis [0029]. A copy of the Risch paper is included herewith.

Thus the use of thousands of bi-allelic covering markers as recited in claims 14 and 24 is supported and is definite. Even if the process of claim 14 used, for example, 2 million covering markers, it would necessarily use thousands of covering markers and be included in the scope of the claim. And even if there were for example, 2 million covering markers in the group of two or more bi-allelic covering markers as recited in claim 24, there would necessarily be thousands of covering markers in the group. And such an embodiment would be within the scope of claim 24.

Regarding new claim 25 Newly added claim 25 contains the language "nearly identical" which caused the Examiner to reject claim 10 for lack of definiteness. Newly added claim 25 deals with redundancy of markers and makes use of description recited in [0315], [0316] and [0317]. Similar description is found in [0268] to [0271]. Applicants respectfully submit that new claim 25 is definite. Specifically when markers are redundant and are in extreme positive linkage disequilibrium then every chromosome in the population that carries allele A also caries allele B and every chromosome that carries not allele A also carries not allele B, or this is nearly the situation [0316]. Under these circumstances the genotype of an individual at one marker will almost always predict the genotype of the individual at the other marker. Similarly allele frequency for an allele at one marker for a sample will predict with a very high degree of certainty or precision the allele frequency for an allele at the other marker.

Though there is relative language in claim 25. This relative language does not render the claim indefinite. As stated in the MPEP 2173.05(b) *"The fact that claim language, including terms of degree, may not be precise, does not automatically render the claim indefinite under 35 USC 112, second paragraph. (Seattle Box Co. v. Industrial Crating & Packing, Inc. 221 USPQ 568). Acceptability of the claim language depends on whether one of ordinary skill in the art would understand what is claimed in light of the specification."* Applicants respectfully submit that the language used in claim 25 is as precise or accurate as the subject matter allows. It is not possible to reasonably specify parameters such as allele frequency differences or departures from maximal linkage disequilibrium between redundant markers with a greater degree of accuracy or precision. A clear result or effect of redundancy of markers (and the relative difference of "nearly identical information") is specified in the claim, specifically that there would be no increase in the likelihood of detecting linkage. The limitation is as follows*: "wherein the inclusion of a bi-allelic marker in the subset so that there would be a redundant pair in the subset would not increase the likelihood of detecting linkage and association of the trait-causing polymorphism".* Applicants respectfully submit the situation is similar to that in Orthokinetics, Inc. v. Safety Chairs, Inc. (1 USPQ 2d 1081), which is cited in MPEP 2173.05(b). In that case, the Court found that a claim to a chair "so dimensioned" as to fit between an automobile doorframe and one of the seats was definite. The Court said the phrase "so dimensioned" is as accurate or precise as the subject matter permits.

Similarly applicants respectfully submit that a redundant marker pair is defined in terms of what it does. The markers of the pair provide nearly identical information, and so the addition of one of the markers does not increase the likelihood of detecting linkage. This is similar to a functional limitation as described in MPEP 2173.05(g), which were found definite in In re Barr (170 USPQ 33) and In re Venezia (189 USPQ 149).

Claims 12 and 22 have also been amended to bring them into harmony with the language in claims from which they depend. Their scope is unchanged.

Conclusion

An RCE has been filed and claims 10, 20 and 24 have been amended in response to the Examiner's rejection and new claim 25 has been added. Claims 12 and 22 have also been amended. Remarks/Arguments in this Response have addressed each point of rejection in the Final Office Action.

For the reasons advanced above, applicants respectfully submit that the application is now in condition for allowance and that action is earnestly solicited.

Respectfully submitted,

Robert O. McGinnis
Registration No. 44, 232

September 13, 2005
1575 West Kagy Blvd.
Bozeman, MT. 59715
tel (406)-522-9355

The Future of Genetic Studies of Complex Human Diseases

Neil Risch and Kathleen Merikangas

Geneticists have made substantial progress in identifying the genetic basis of many human diseases, at least those with conspicuous determinants. These successes include Huntington's disease, Alzheimer's disease, and some forms of breast cancer. However, the detection of genetic factors for complex diseases—such as schizophrenia, bipolar disorder, and diabetes—has been far more complicated. There have been numerous reports of genes or loci that might underlie these disorders, but few of these findings have been replicated. The modest nature of the gene effects for these disorders likely explains the contradictory and inconclusive claims about their identification. Despite the small effects of such genes, the magnitude of their attributable risk (the proportion of people affected due to them) may be large because they are quite frequent in the population, making them of public health significance.

Has the genetic study of complex disorders reached its limits? The persistent lack of replicability of these reports of linkage between various loci and complex diseases might imply that it has. We argue below that the method that has been used successfully (linkage analysis) to find major genes has limited power to detect genes of modest effect, but that a different approach (association studies) that utilizes candidate genes has far greater power, even if one needs to test every gene in the genome. Thus, the future of the genetics of complex diseases is likely to require large-scale testing by association analysis.

How large does a gene effect need to be in order to be detectable by linkage analysis? We consider the following model: Suppose a disease susceptibility locus has two alleles A and a, with population frequencies p and $q = 1 - p$, respectively. There are three genotypes: AA, Aa, and aa. We define genotypic relative risks (GRR, the increased chance that an individual with a particular genotype has the disease) as follows: Let the risk for individuals of genotype Aa be γ times greater than the risk for individuals with genotype aa, a GRR of γ. We assume a multiplicative relation for two A alleles, so that the GRR for genotype AA is γ^2. The method of linkage analysis we have chosen for this argument is a popular current paradigm in which pairs of siblings, both with the disease, are examined for sharing of alleles at multiple sites in the genome defined by genetic markers. The more often the affected siblings share the same allele at a particular site, the more likely the site is close to the disease gene. Using the formulas in (*1*), we calculate the expected proportion Y of alleles shared by a pair of affected siblings for the best possible case—that is, a closely linked marker locus (recombination fraction $\theta = 0$) that is fully informative (heterozygosity = 1) (*2*)—as

$$Y = \frac{1 + w}{2 + w} \text{ where } w = \frac{pq\,(\gamma - 1)^2}{(p\gamma + q)^2}$$

If there is no linkage of a marker at a particular site to the disease; the siblings would be expected to share alleles 50% of the time; that is, Y would equal 0.5. Values of Y for various values of p and γ are given in the third column of the table. For an allele of moderate frequency (p is 0.1 to 0.5) that confers a GRR (γ) of fourfold or greater, there is a detectable deviation of Y from the null value of 0.5. On the other hand, for an allele conferring a GRR of 2 or less, the expected marker-sharing only marginally exceeds 50%, for any allele frequency (p). Thus, it is clear that the use of linkage analysis for loci conferring GRR of about 2 or less will never allow identification because the number of families required (more than ~2500) is not practically achievable.

Although tests of linkage for genes of modest effect are of low power, as shown by the above example, direct tests of association with a disease locus itself can still be quite strong. To illustrate this point, we use the transmission/disequilibrium test of Spielman *et al.* (*3*). In this test, transmission of a particular allele at a locus from heterozygous parents to their affected offspring is examined. Under Mendelian inheritance, all alleles should have a 50% chance of being transmitted to the next generation. In contrast, if one of the alleles is associated with disease risk, it will be transmitted more often than 50% of the time.

For this approach, we do not need families with multiple affected siblings, but can focus just on single affected individuals and their parents. For the same model given above, we can calculate the proportion of heterozygous parents as $pq(\gamma + 1)/(p\gamma + q)$(*4*). Similarly, the probability for a heterozygote parent to transmit the high risk A allele is just $\gamma/(1 + \gamma)$. Association tests can also be performed for pairs of affected siblings. When the locus is associated with disease, the transmission excess over 50% is the same as for single offspring, but the probability of parental heterozygosity is increased at low values of p; for higher values of p, the probability of parental heterozygosity is decreased. The formula for parental heterozygosity for an affected pair of siblings for the same genetic model as used in the first example is

$$h = \frac{pq\,(\gamma + 1)^2}{2(p\gamma + q)^2 + pq(\gamma - 1)^2}$$

		Linkage			Association			
					Singletons		Sib pairs	
Genotypic risk ratio (γ)	Frequency of disease allele A (p)	Probability of allele sharing (Y)	No. of families required (N)	Probability of transmitting disease allele A P(tr-A)	Proportion of heterozygous parents (Het)	(N)	(Het)	(N)
4.0	0.01	0.520	4260	0.800	0.048	1098	0.112	235
	0.10	0.597	185	0.800	0.346	150	0.537	48
	0.50	0.576	297	0.800	0.500	103	0.424	61
	0.80	0.529	2013	0.800	0.235	222	0.163	161
2.0	0.01	0.502	296,710	0.667	0.029	5823	0.043	1970
	0.10	0.518	5382	0.667	0.245	695	0.323	264
	0.50	0.526	2498	0.667	0.500	340	0.474	180
	0.80	0.512	11,917	0.667	0.267	640	0.217	394
1.5	0.01	0.501	4,620,807	0.600	0.025	19,320	0.031	7776
	0.10	0.505	67,816	0.600	0.197	2218	0.253	941
	0.50	0.510	17,997	0.600	0.500	949	0.490	484
	0.80	0.505	67,816	0.600	0.286	1663	0.253	941

Comparison of linkage and association studies. Number of families needed for identification of a disease gene.

N. Risch is in the Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305–5120, USA. E-mail: risch@lahmed.stanford.edu. K. Merikangas is in the Departments of Epidemiology and Psychiatry, Unit, Yale University School of Medicine, New Haven, CT 06510, USA. E-mail: kath@zeus.psych.yale.edu

On the right side of the table, we present the proportion of heterozygous parents (Het) and the probability of transmission of the A allele from a heterozygous parent to an affected child [P(tr-A)] for the same values of GRR as considered above for the example of linkage analysis. The deviation from the null hypothesis of 50% transmission from heterozygous parents is substantially greater than the excess allele sharing that is found by linkage analysis in sibling pairs. This disparity between the methods is particularly true for lower values of γ (that is, with lower relative risk). For example, for $\gamma = 1.5$, allele sharing is at most 51%, while the A allele is transmitted 60% of the time from heterozygous parents.

In this respect then, association studies seem to be of greater power than linkage studies. But of course, the limitation of association studies is that the actual gene or genes involved in the disease must be tentatively identified before the test can be performed. In fact, the actual polymorphism within the gene (or at least a polymorphism in strong disequilibrium) must be available. However, we show that this requirement is only daunting because of limitations imposed by current technological capabilities, not because sufficient families with the disease are not available or the statistical power is inadequate (*5*). For example, imagine the time when all human genes (say 100,000 in total) have been found and that simple, diallelic polymorphisms in these genes have been identified. Assume that five such diallelic polymorphisms have been identified within each gene, so that a total of $10 \times 10^5 = 10^6$ alleles need to be tested. The statistical problem is that the large number of tests that need to be made leads to an inflation of the type 1 error probability. For a linkage test with pairs of affected siblings, we use a lod score (logarithm of the odds ratio for linkage) criterion of 3.0, which asymptotically corresponds to a type 1 error probability α of about 10^{-4}. In a linkage genome screen with 500 markers, this significance level gives a probability greater than 95% of no false positives. The equivalent false positive rate for 1,000,000 independent association tests can be obtained with a significance level $\alpha = 5 \times 10^{-8}$.

We illustrate the power of linkage versus association tests at different significance levels by determining the sample size N (number of families) necessary to obtain 80% power (the probability of rejecting the null hypothesis when it is false) (*6*) (see table). With a linkage approach and a disease gene with a GRR of 4 or greater, the number of affected sibling pairs necessary to detect linkage is realistic (185 or 297), provided the allele frequency p is between 5 and 75%. For a gene with a GRR of 2 or less, however, the sample sizes are generally beyond reach (well over 2000), precluding their identification by this approach. In contrast, the required sample size for the association test, even allowing for the smaller significance level, is vastly less than for linkage, especially for affected sibling pair families when the value of p is small. Even for a GRR of 1.5, the sample sizes are generally less than 1000, well within reason.

Thus, the primary limitation of genome-wide association tests is not a statistical one but a technological one. A large number of genes (up to 100,000) and polymorphisms (preferentially ones that create alterations in derived proteins or their expression) must first be identified, and an extremely large number of such polymorphisms will need to be tested. Although testing such a large number of polymorphisms on several hundred, or even a thousand families, might currently seem implausible in scope, more efficient methods of screening a large number of polymorphisms (for example, sample pooling) may be possible. Furthermore, the number of tests we have used as the basis for our calculations (1,000,000) is likely to be far larger than necessary if one allows for linkage disequilibrium, which could substantially reduce the required number of markers and families needed for initial screening.

Some of the important loci for complex diseases will undoubtedly be found by linkage analysis. However, the limitations to detecting many of the remaining genes by linkage studies can be overcome; numerous genetic effects too weak to identify by linkage can be detected by genomic association studies. Fortunately, the samples currently collected for linkage studies (for example, affected pairs of siblings and their parents) can also be used for such association studies. Thus, investigators should preserve their samples for future large-scale testing.

The human genome project can have more than one reward. In addition to sequencing the entire human genome, it can lead to identification of polymorphisms for all the genes in the human genome and the diseases to which they contribute. It is a charge to the molecular technologists to develop the tools to meet this challenge and provide the information necessary to identify the genetic basis of complex human diseases.

References and Notes

1. N. Risch, *Am. J. Hum. Genet.* **40**, 1 (1987); *ibid.* **46**, 229 (1990).
2. From the formulas in (*1*), we have $\lambda_O = 1 + 0.5V_A/K^2$ and $\lambda_S = 1 + (0.5V_A + 0.25V_D)/K^2$, where $K = p^2\gamma^2 + 2pq\gamma + q^2 = (p\gamma + q)^2$, $V_A = 2pq(\gamma - 1)^2/(p\gamma + q)^2$, and $V_D = p^2q^2(\gamma - 1)^2$. Hence, $\lambda_O = 1 + w$ and $\lambda_S = (1 + 0.5w)^2$, where $w = pq(\gamma - 1)^2$. The proportion of alleles shared is given by $Y = 1 - 0.5z_1 - z_0$, where z_1 and z_0 are the probabilities of the sib pair sharing 1 and 0 disease alleles ibd, respectively. From (*1*), $z_0 = 0.25/\lambda_S$ and $z_1 = 0.5\lambda_O/\lambda_S$. Thus, after some algebra, $Y = 1 - 0.25(\lambda_O + 1)/\lambda_S = (1 + w)/(2 + w)$.
3. R. Spielman, R. E. McGinnis, W. J. Ewens, *Am. J. Hum. Genet.* **52**, 506 (1993).
4. By Bayes theorem, the probability of a parent of an affected child being heterozygous is given by P(HetlAff child) = P(Het)P(Aff ChildlHet)/P(Aff Child) = $2pq[0.5p(\gamma^2 + \gamma)] + 0.5q(\gamma + 1)/(p\gamma + q)^2 = pq(\gamma + 1)/(p\gamma + q)$.
5. E. S. Lander and N. J. Schork, *Science* **265**, 2037 (1994).
6. Consider a set of M independent, identically distributed random variables B_i of discrete value. Under the null hypothesis H_0, assume $E(B_i) = 0$ and $\mathrm{Var}(B_i) = 1$. Under the alternative hypothesis H_1, let $E(B_i) = \mu$ and $\mathrm{Var}(B_i) = \sigma^2$. For a sample of size M, let $T = \Sigma B_i/\sqrt{M}$. Then under H_0, T also has mean 0 and variance 1, while under H_1, it has mean $\sqrt{M}\mu$ and variance σ^2. We assume that T is approximately normally distributed both under H_0 and H_1. Then the sample size M required to obtain a power of $1 - \beta$ for a significance level α is given by

$$M = (Z_\alpha - \sigma Z_{1-\beta})^2/\mu^2 \quad (1)$$

For each affected sib pair, we score the number of alleles shared ibd from each of $2N$ parents. Define $B_i = 1$ if an allele is shared from the ith parent and $B_i = -1$ if unshared. Under the null hypothesis of no linkage, $P(B_i = 1) = P(B_i = -1) = 0.5$, so $E(B_i) = 0$ and $\mathrm{Var}(B_i) = 1$. For the genetic model described above with genotypic relative risks of γ^2, γ, and 1, allele sharing by affected sibs is independent for the two parents; thus, we can consider sharing of alleles one parent at a time. Thus, for affected sib pairs assuming $\theta = 0$ and no linkage disequilibrium, the formula is

$$N = \frac{(Z_\alpha - \sigma Z_{1-\beta})^2}{2\mu^2}$$

where

$$\mu = 2Y - 1$$

$$\sigma^2 = 4Y(1 - Y)$$

$$Y = \frac{1 + w}{2 + w}$$

$$w = \frac{pq(\gamma - 1)^2}{(p\gamma + q)^2}$$

$Z_\alpha = 3.72$ (corresponding to $\alpha = 10^{-4}$), and $Z_{1-\beta} = -0.84$ (corresponding to $1 - \beta = 0.80$). For an association test using the transmission/disequilibrium test, with the disease locus or a nearby locus in complete disequilibrium, the number (N) of families with affected singletons required for 80% power is also calculated from formula 1. For this case, we score the number of transmissions of allele A from heterozygous parents. Let h be the probability a parent is heterozygous under the alternative hypothesis, namely, $h = pq(\gamma + 1)/(p\gamma + q)$. Then define $B_i = h^{-0.5}$ if the parent is heterozygous and allele A is transmitted; $B_i = 0$ if the parent is homozygous; and $B_i = -h^{-0.5}$ if the parent is heterozygous and transmits allele a. Under the null hypothesis, $E(B_i) = 0$ and $\mathrm{Var}(B_i) = 1$. Under the alternative hypothesis, $\mu = E(B_i) = \sqrt{h}(\gamma - 1)/(\gamma + 1)$ and $\sigma^2 = \mathrm{Var}(B_i) = 1 - h(\gamma - 1)^2/(\gamma + 1)^2$. In this case, there are two parents per family and they act independently, so the required number (N) of families is given by half of formula 1 where μ and σ^2 are given above. Here, $Z_\alpha = 5.33$ (corresponding to $\alpha = 5 \times 10^{-8}$). For the same test but with affected sib pairs instead of singletons, the number of families required is given by half of formula 1 (transmissions from two parents to two children) with the same formulas for μ and σ^2 as for singleton families but now using the heterozygote frequency for parents of affected sib pairs. Using the above formulas, we can calculate sample sizes for the three study designs.

27 October 1995; accepted 6 June 1996.

Proc. Natl. Acad. Sci. USA
Vol. 94, pp. 13057–13062, November 1997
Genetics

Yeast microarrays for genome wide parallel genetic and gene expression analysis

DEVAL A. LASHKARI*†, JOSEPH L. DERISI‡, JOHN H. MCCUSKER§, ALLEN F. NAMATH‡, CRISTL GENTILE§, SEUNG Y. HWANG‡, PATRICK O. BROWN‡, AND RONALD W. DAVIS*‡

Departments of *Genetics and ‡Biochemistry, Stanford University, Stanford, CA 94305; and §Department of Microbiology, Duke University, Durham, NC 27710

Contributed by Ronald W. Davis, September 2, 1997 ORF = Open reading frame

ABSTRACT **We have developed high-density DNA microarrays of yeast ORFs. These microarrays can monitor hybridization to ORFs for applications such as quantitative differential gene expression analysis and screening for sequence polymorphisms. Automated scripts retrieved sequence information from public databases to locate predicted ORFs and select appropriate primers for amplification. The primers were used to amplify yeast ORFs in 96-well plates, and the resulting products were arrayed using an automated micro arraying device. Arrays containing up to 2,479 yeast ORFs were printed on a single slide. The hybridization of fluorescently labeled samples to the array were detected and quantitated with a laser confocal scanning microscope. Applications of the microarrays are shown for genetic and gene expression analysis at the whole genome level.**

The genome sequencing projects have generated and will continue to generate enormous amounts of sequence data. The genomes of *Saccharomyces cerevisiae, Haemophilus influenzae* (1), *Mycoplasma genitalium* (2), and *Methanococcus jannischii* (3) have been completely sequenced. Other model organisms have had substantial portions of their genomes sequenced as well including the nematode *Caenorhabditis elegans* (4) and the small flowering plant *Arabidopsis thaliana* (5). Given this ever-increasing amount of sequence information, new strategies are necessary to efficiently pursue the next phase of the genome projects—the elucidation of gene expression patterns and gene product function on a whole genome scale.

One important use of genome sequence data is to attempt to identify the functions of predicted ORFs within the genome. Many of the ORFs identified in the yeast genome sequence were not identified in decades of genetic studies and have no significant homology to previously identified sequences in the database. In addition, even in cases where ORFs have significant homology to sequences in the database, or have known sequence motifs (e.g., protein kinase), this is not sufficient to determine the actual biological role of the gene product. Experimental analysis must be performed to thoroughly understand the biological function of a given ORF's product. Model organisms, such as *S. cerevisiae*, will be extremely important in improving our understanding of other more complex and less manipulable organisms.

To examine in detail the functional role of individual ORFs and relationships between genes at the expression level, this work describes the use of genome sequence information to study large numbers of genes efficiently and systematically. The procedure was as follows. (*i*) Software scripts scanned annotated sequence information from public databases for predicted ORFs. (*ii*) The start and stop position of each identified ORF was extracted automatically, along with the sequence data of the ORF and 200 bases flanking either side. (*iii*) These data were used to automatically select PCR primers that would amplify the ORF. (*iv*) The primer sequences were automatically input into the automated multiplex oligonucleotide synthesizer (6). (*v*) The oligonucleotides were synthesized in 96-well format, and (*vi*) used in 96-well format to amplify the desired ORFs from a genomic DNA template. (*vii*) The products were arrayed using a high-density DNA arrayer (7–10). The gene arrays can be used for hybridization with a variety of labeled products such as cDNA for gene expression analysis or genomic DNA for strain comparisons, and genomic mismatch scanning purified DNA for genotyping (11).

METHODS

Script Design. All scripts were written in UNIX Tool Command Language. Annotated sequence information from GenBank was extracted into one file containing the complete nucleotide sequence of a single chromosome. A second file contained the assigned ORF name followed by the start and stop positions of that ORF. The actual sequence contained within the specified range, along with 200 bases of sequence flanking both sides, was extracted and input into the primer selection program PRIMER 0.5 (Whitehead Institute, Boston). Primers were designed so as to allow amplification of entire ORFs. The selected primer sequences were read by the 96-well automated multiplex oligonucleotide synthesizer instrument for primer synthesis. The forward and reverse primers were synthesized in two separate 96-well plates in corresponding wells. All primers were synthesized on a 20-nmol scale.

ORF Amplification and Purification. Genomic DNA was isolated as described (12) and used as template for the amplification reactions. Each PCR was done in a total volume of 100 μl. A total of 0.2 μM each of forward and reverse primers were aliquoted into a 96-well PCR plate (Robbins Scientific, Sunnyvale, CA); a master mix containing 0.24 mM each dNTP, 10 mM Tris (pH 8.5), 50 mM $MgCl_2$, 2.5 units *Taq* polymerase, and 10 ng of template was added to the primers, and the entire mix was thermal cycled for 30 cycles as follows: 15 min at 94°C, 15 min at 54°C, and 30 min at 72°C. Products were ethanol precipitated in polystyrene v-bottom 96-well plates (Costar). All samples were dried and stored at −20°C.

Arraying Procedure and Processing. Microarrays were made as described (8).

A custom built arraying robot was used to print batches of 48 slides. The robot utilizes four printing tips which simultaneously pick up ≈1 μl of solution from 96-well microtiter plates. After printing, the microarrays were rehydrated for 30 sec in a humid chamber and then snap dried for 2 sec on a hot plate (100°C). The DNA was then UV crosslinked to the surface by subjecting the slides to 60 millijoules of energy. The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reaction, the bound DNA was denatured by a 2-min incubation in distilled water at ≈95°C.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: YEP, yeast extract/peptone.

†To whom reprint requests should be sent at the present address: Synteni, Inc., 6519 Dumbarton Circle, Fremont, CA 94555.

DISCUSSION

The results of these experiments show that many genes are differentially expressed under the three environmental conditions described here. The expected and predicted changes in gene expression, such as HSP12 in the heat-shocked culture, TIP1 in the cold-shocked culture, and GAL2 in the steady-state galactose culture, were observed in every case. However, in addition to the expected changes in gene expression, significant differential expression was also observed for many other genes that would not, *a priori*, be expected to be differentially expressed. For example, expression of PHO11 decreased and expression of YLR194, KIN2, and HXT6 increased in the heat shocked culture. Expression of MST1 and APE3 decreased and expression of PDR5 and GAR1 increased in the cold-shocked culture. In addition, ADE4 and SER2 were expressed at reduced levels whereas PHO84 and ACH1 were expressed at higher levels in cells grown in galactose compared with cells grown in glucose. Differential expression of these and many other genes was specific to one of these three environmental conditions.

Many other genes were found to be differentially expressed under more than one condition. When differentially expressed genes in cold- and heat-shocked cultures were compared, 30 genes were found in common. Of these 30 genes, 28 showed inverse expression (i.e., increased expression under one condition and decreased expression under the other condition). Two genes, YCR058 and YKL102, showed elevated expression in response to both cold and heat shock. Fifteen genes were found to be differentially expressed in both the heat-shocked and steady-state galactose cultures: 9 genes showed increased expression and 5 showed decreased expression under both conditions. Twenty genes were differentially expressed in both the cold-shocked and steady-state galactose cultures: 8 genes showed decreased expression and 5 genes showed increased expression under both conditions. Six genes showed increased expression in the galactose culture and decreased expression in the cold shocked culture. One gene (ODP1) showed increased expression in both the cold-shocked and steady-state galactose cultures.

Gene expression is affected in a global fashion when environmental conditions are changed and both expected and unexpected genes are affected. There is also overlap in the genes that are differentially expressed under quite different environmental conditions. These results can be rationalized by considering the high degree of cross-pathway regulation in yeast. For example, there is evidence for cross-pathway regulation between (*i*) carbon and nitrogen metabolism (18), (*ii*) phosphate and sulfate metabolism (19), and (*iii*) purine, phosphate, and amino acid metabolism (20–24). There are also examples of the interaction of general and specific transcription factors (25, 26). Finally, within the broad class of amino acid biosynthetic genes, there is evidence for amino acid specific regulation of some genes, regulation via general control for other genes, and regulation via both specific and general control for other genes (22, 27–30).

Cross-pathway regulation arises from the complex structure of promoters. Virtually all promoters contain sites for multiple transcription factors and, therefore, virtually all genes are subject to combinatorial regulation. For example, the HIS4 promoter contains binding sites for GCN4 (the general amino acid control transcription factor), PHO2/BAS2 (a transcriptional regulator of phosphatase and purine biosynthetic genes), and BAS1 (a transcriptional regulator of purine biosynthetic genes) (31). It is likely that the complex effects on gene expression described in this work are a direct consequence of the combinatorial regulation of gene expression.

These findings illustrate the power of the highly parallel whole genome approach when examining gene expression. The global effects of environmental change on gene expression can now be directly visualized. It is clear that determining the mechanism(s) and the functional role of the dramatic global effects on gene expression in different environments will be a significant challenge. The era of whole genome analysis will, ultimately, allow researchers to switch from the very focused single gene/promoter view of gene expression and instead view the cell more as a large complex network of gene regulatory pathways.

With the entire sequence of this model organism known, new approaches have been developed that allow for genome wide analyses (32, 33) of gene function. The genome microarrays represent a novel tool for genetic and expression analysis of the yeast genome. This pilot study uses arrays containing >35% of the yeast ORFs and it is clear that the entire set of ORFs from the yeast genome can be arrayed using the directed primer based strategy detailed here. Recent advances in arraying technology will allow all 6,100 ORFs to be arrayed in an area of less than 1.8 cm^2. Furthermore, as the technology improves, detection limits will allow less than 500 ng of starting mRNA material to be used for making probe.

The genome arrays provide for a robust, fully automated approach toward examining genome structure and gene function. They allow for comparisons between different genomes as well as a detailed study of gene expression at the global level. This research will help to elucidate relationships between genes and allow the researcher to understand gene function by understanding expression patterns across the yeast genome.

Support was provided by National Institutes of Health Grant P0/HG00205.

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* 269, 496–512.
2. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* 270, 397–403.
3. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., *et al.* (1996) *Science* 273, 1058–1073.
4. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., *et al.* (1992) *Nature (London)* 356, 37.
5. Newman, T., de Bruijn, F. J., Green, P., Keegstra, K., Kende, H., *et al.* (1994) *Plant Physiol.* 106, 1241–1255.
6. Lashkari, D. A., Hunicke-Smith, S. P., Norgren, R. M., Davis, R. W. & Brennan, T. (1995) *Proc. Natl. Acad. Sci. USA* 92, 7912–7915.
7. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* 270, 467–470.
8. Shalon, D., Smith, S. & Brown, P. O. (1996) *Genome Res.* 6, 639–645.
9. Heller, R. A., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., Woolley, D. E. & Davis, R. W. (1997) *Proc. Natl. Acad. Sci. USA* 94, 2150–2155.
10. DeRisi, J., Penland, L., brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su Ya & Trent, J. M. (1996) *Nat. Genet.* 14, 457–460.
11. Nelson, S. F., McCusker, J. H., Sander, M. A., Kee, Y., Modrich, P. & Brown P. O. (1993) *Nat. Genet.* 4, 11–18.
12. Hoffman, C. S. & Winston, F. (1989) *Gene* 84, 473–479.
13. Schmitt, M., Brown, T. & Trumpower, B. (1990) *Nucleic Acids Res.* 18, 3091.
14. Ehrenhofer-Murray, A. E., Wurgler, F. E. & Sengstag, C. (1994) *Mol. Gen. Genet.* 244, 287–294.
15. Kim, K-W., Kamerud, J. Q., Livingston, D. M. & Roon, R. J. (1988) *J. Biol. Chem.* 263, 11948–11953.
16. Kim, K.-W. & Roon, R. J. (1984) *J. Bacteriol.* 157, 958–961.
17. Craig, E. A. (1992) in *The Molecular Biology of the Yeast Saccharomyces: Gene Expression*, eds. Jones, E. W., Pringle, J. R. & Broach, J. R. (Cold Spring Harbor Lab. Press, Plainview, NY), Vol. 2, pp. 501–537.
18. Dang, V. D., Bohn, C., Bolotin-Fukuhara, M. & Daignan-Fornier, B. (1996) *J. Bacteriol.* 178, 1842–1849.
19. O'Connell, K. F. & Baker, R. E. (1992) *Genetics* 132, 63–73.
20. Braus, G., Mosch, H. U., Vogel, K., Hinnen, A. & Hutter, R. (1989) *EMBO J.* 8, 939–945.
21. Mosch, H. U., Scheier, B., Lahti, R., Mantsala, P. & Braus, G. H. (1991) *J. Biol. Chem.* 266, 20453–20456.
22. Mitchell, A. P. & Magasanik, B. (1984) *Mol. Cell. Biol.* 4, 2767–2773.
23. Daignan-Fornier, B. & Fink, G. R. (1992) *Proc. Natl. Acad. Sci USA* 89, 6746–6750.
24. Tice-Baldwin, K., Fink, G. R. & Arndt, K. T. (1989) *Science* 246, 931–935.
25. Messenguy, F. & Dubois, E. (1993) *Mol. Cell. Biol.* 13, 2586–2592.
26. Devlin, C., Tice-Baldwin, K., Shore, D. & Arndt, K. T. (1991) *Mol. Cell. Biol.* 11, 3642–3651.
27. Magasanik, B. (1992) in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression*, eds. Jones, E. W., Pringle, J. R. & Broach, J. R. (Cold Spring Harbor Lab. Press, Plainview, NY), Vol. 2, pp. 283–317.
28. Hinnebusch, A. G. (1992) in *The Molecular and Cellular Biology of the Yeast Saccharomyces: Gene Expression*, eds. Jones, E. W., Pringle, J. R. & Broach, J. R. (Cold Spring Harbor Lab. Press, Plainview, NY), Vol. 2, pp. 319–414.
29. Brisco, P. R. & Kohlhaw, G. B. (1990) *J. Biol. Chem.* 265, 11667–11675.
30. O'Connell, K. F., Surdin-Kerjan, Y. & Baker R. E. (1995) *Mol. Cell. Biol.* 15, 1879–1888.
31. Arndt K. T., Styles, C. & Fink, G. R. (1987) *Science* 237, 874–880.
32. Smith, V., Chou, K. N., Lashkari, D., Botstein, D. & Brown, P. O. (1996) *Science* 274, 2069–2074.
33. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittman, M. & Davis, R. W. (1996) *Nat. Genet.* 14, 450–456.

D–0.1 M KCl, Tat-SF/pp140 was eluted with increasing salt concentrations and was detected mostly in 0.2 to 0.4 M KCl fractions. These fractions were pooled, dialyzed against buffer D–0.1 M KCl, and loaded onto a glutathione Sepharose (Pharmacia) column containing GST-Tat fusion proteins. After the column was washed with buffer D–0.4 M KCl, Tat-SF/pp140 was eluted from the column with buffer D containing 1.4 M KCl. The estimated overall purification after these steps was ~3000-fold. In the experiment shown in Fig. 3, the 0.2 to 0.4 M KCl heparin Sepharose fraction containing Tat-SF activity was subjected to fractionation through an Affi-Gel 10 matrix column (Bio-Rad) containing immobilized Tat. Tat-SF activity was eluted from the column with increasing salt concentrations. The 0.6 M KCl fraction was analyzed as described in Fig. 3.

10. T. O'Brien, S. Hardin, A. Greenleaf, J. T. Lis, *Nature* **370**, 75 (1994); M. E. Dahmus, *Biochim. Biophys. Acta.* **1261**, 171 (1995).
11. A. P. Rice and F. Carlotti, *J. Virol.* **64**, 1864 (1990).
12. The Tat-SF/pp140 fraction eluted from the GST-Tat column was subjected to SDS–polyacrylamide gel electrophoresis (PAGE), and the pp140 polypeptide was blotted onto a nitrocellulose membrane. Approximately 15 μg of pp140 were recovered from the membrane and subjected to digestion with lys-C. Six major peptides were obtained and microsequenced. One of the peptides (KMNAQETATGMAFEEPIDE) was contained in the sequence of EST60354 in the Washington University–Merck EST database. An Xho I–Eco RI fragment corresponding to the COOH-terminus of the Tat-SF1 gene and its 3′ untranslated region was labeled and used as a probe to screen a λZipLox (Gibco BRL) cDNA library prepared from human HL60 cells. Complementary DNAs were recovered from seven independent plaques in the autonomously replicating plasmid pZL1 as instructed by the manufacturer (Gibco BRL). The largest cDNA clone containing the full-length Tat-SF1 gene was named pZL-Tat-SF1-4b and was sequenced by dideoxy-DNA sequencing with T7 DNA polymerase.
13. D. R. Marshak and D. Carroll, *Methods Enzymol.* **200**, 134 (1991).
14. D. J. Kenan, C. C. Query, J. D. Keene, *Trends Biochem. Sci.* **16**, 214 (1991).
15. O. Delattre *et al.*, *Nature* **359**, 162 (1992); P. H. Sorensen *et al.*, *Nature Genet.* **6**, 146 (1994).
16. A. Crozat, P. Aman, N. Mandahl, D. Ron, *Nature* **363**, 640 (1993); T. H. Rabbitts, A. Forster, R. Larson, P. Nathan, *Nature Genet.* **4**, 175 (1993).
17. M. Ladanyi, *Diagn. Mol. Pathol.* **4**, 162 (1995); T. H. Rabbitts, *Nature* **372**, 143 (1994).
18. S. E. Harper, Y. Qiu, P. A. Sharp, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 8536 (1996).
19. J. W. Lillie and M. R. Green, *Nature* **338**, 39 (1989).
20. H. Kato *et al.*, *Genes Dev.* **6**, 655 (1992); R. A. Marciniak and P. A. Sharp, *EMBO J.* **10**, 4189 (1991).
21. M. G. Izban and D. S. Luse, *Genes Dev.* **6**, 1342 (1992); D. Wang and D. K. Hawley, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 843 (1993).
22. E. Bengal, O. Flores, A. Krauskopf, D. Reinberg, Y. Aloni, *Mol. Cell. Biol.* **11**, 1195 (1991); J. Greenblatt, J. R. Nodwell, S. W. Mason, *Nature* **364**, 401 (1993).
23. C. H. Herrmann and A. P. Rice, *J. Virol.* **69**, 1612 (1995).
24. N. A. McMillan *et al.*, *Virology* **213**, 413 (1995).
25. W. A. May *et al.*, *Mol. Cell. Biol.* **13**, 7393 (1993); H. Zinszner, R. Albalat, D. Ron, *Genes Dev.* **8**, 2513 (1994); D. D. Prasad, M. Ouchida, L. Lee, V. N. Rao, E. S. Reddy, *Oncogene* **9**, 3717 (1994).
26. P. J. Mitchell and R. Tjian, *Science* **245**, 371 (1989).
27. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
28. M. A. Truett *et al.*, *DNA* **4**, 333 (1985).
29. H. E. Gendelman *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 9759 (1986).
30. L. S. Tiley, P. H. Brown, B. R. Cullen, *Virology* **178**, 560 (1990).
31. J. R. Neumann, C. A. Morency, K. O. Russian, *BioTechniques* **5**, 444 (1987).

32. We are grateful to B. Pepinsky and Biogen for providing pure HIV Tat protein and Tat mutant TatΔC; to J. Borrow [Massachusetts Institute of Technology (MIT) Center for Cancer Research] for human cDNA libraries; and to R. Cook (MIT Biopolymers Laboratory) for peptide sequencing. We thank K. Luo, J. Borrow, and H. Kawasaki for valuable advice and discussions; and B. Blencowe, K. Cepek, G. Jones, K. Luo, and C. Query for helpful comments on the manuscript. We also thank M. Siafaca for secretarial support. Supported by grants from the National Institutes of Health (GM34277 and AI32486) to P.A.S., and partially supported by a National Cancer Institute Center core grant (CA14051), was supported by a postdoctoral fellowship of The Coffin Childs Memorial Fund for Medical Research.

19 June 1996; accepted 23 August 1996

Accessing Genetic Information with High-Density DNA Arrays

Mark Chee, Robert Yang, Earl Hubbell, Anthony Berno, Xiaohua C. Huang, David Stern, Jim Winkler, David J. Lockhart, Macdonald S. Morris, Stephen P. A. Fodor

Rapid access to genetic information is central to the revolution taking place in molecular genetics. The simultaneous analysis of the entire human mitochondrial genome is described here. DNA arrays containing up to 135,000 probes complementary to the 16.6-kilobase human mitochondrial genome were generated by light-directed chemical synthesis. A two-color labeling scheme was developed that allows simultaneous comparison of a polymorphic target to a reference DNA or RNA. Complete hybridization patterns were revealed in a matter of minutes. Sequence polymorphisms were detected with single-base resolution and unprecedented efficiency. The methods described are generic and can be used to address a variety of questions in molecular genetics including gene expression, genetic linkage, and genetic variability.

A central theme in modern genetics is the relation between genetic variability and phenotype. To understand genetic variation and its consequences on biological function, an enormous effort in comparative sequence analysis will need to be carried out. Conventional nucleic acid sequencing technologies make use of analytical separation techniques to resolve sequence at the single nucleotide level (*1*, *2*). However, the effort required increases linearly with the amount of sequence. In contrast, biological systems read, store, and modify genetic information by molecular recognition (*3*). Because each DNA strand carries with it the capacity to recognize a uniquely complementary sequence through base pairing, the process of recognition, or hybridization, is highly parallel, as every nucleotide in a large sequence can in principle be queried at the same time. Thus, hybridization can be used to efficiently analyze large amounts of nucleotide sequence. In one proposal, sequences are analyzed by hybridization to a set of oligonucleotides representing all possible subsequences (*4*). A second approach, used here, is hybridization to an array of oligonucleotide probes designed to match specific sequences. In this way the most informative subset of probes is used. Implementation of these concepts relies on recently developed combinatorial technologies to generate any ordered array of a large number of oligonucleotide probes (*5*).

The fundamentals of light-directed oligonucleotide array synthesis have been described (*5*, *6*). Any probe can be synthesized at any discrete, specified location in the array, and any set of probes composed of the four nucleotides can be synthesized in a maximum of $4N$ cycles, where N is the length of the longest probe in the array. For example, the entire set of $\sim 10^{12}$ 20-nucleotide oligomer probes, or any desired subset, can be synthesized in only 80 coupling cycles. The number of different probes that can be synthesized is limited only by the physical size of the array and the achievable lithographic resolution (*7*).

An array consisting of oligonucleotides complementary to subsequences of a target sequence can be used to determine the identity of a target sequence, measure its amount, and detect differences between the target and a reference sequence. Many different arrays can be designed for these purposes. One such design, termed a 4L tiled array, is depicted in Fig. 1A. In each set of four probes, the perfect complement will hybridize more strongly than mismatched probes. By this approach, a nucleic acid target of length L can be scanned for mutations with a tiled array containing $4L$ probes. For example, to query the 16,569 base pairs (bp) of human mitochondrial DNA (mtDNA), only 66,276 probes of the possible $\sim 10^9$ 15-nucleotide oligomers need to be used.

The use of a tiled array of probes to read a target sequence is illustrated in Fig. 1C. A tiled array of 15-nucleotide oligomers varied

Affymetrix, 3380 Central Expressway, Santa Clara, CA 95051, USA.

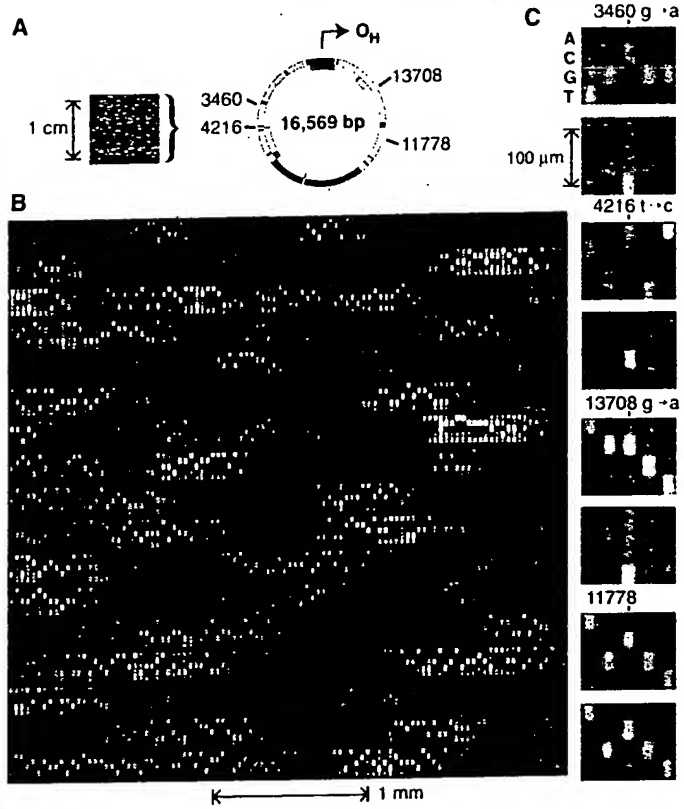


Fig. 3. Human mitochondrial genome on a chip. (**A**) An image of the array hybridized to 16.6 kb of mitochondrial target RNA (L strand). The 16,569-bp map of the genome is shown, and the H strand origin of replication (O_H), located in the control region, is indicated. (**B**) A portion of the hybridization pattern magnified. In each column there are five probes: A, C, G, T, and Δ, from top to bottom. The Δ probe has a single-base deletion instead of a substitution and hence is 24 instead of 25 bases in length. The scale is indicated by the bar beneath the image. Although there is considerable sequence-dependent intensity variation, most of the array can be read directly. The image was collected at a resolution of ~100 pixels per probe cell. (**C**) The ability of the array to detect and read single-base differences in a 16.6-kb sample is illustrated. Two different target sequences were hybridized in parallel to different chips. The hybridization patterns are compared for four different positions in the sequence. Only the $P^{25,13}$ probes are shown. The top panel of each pair shows the hybridization of the mt3 target, which matches the chip P^0 sequence at these positions. The lower panel shows the pattern generated by a sample from a patient with Leber's hereditary optic neuropathy (LHON). Three known pathogenic mutations, LHON3460, LHON4216, and LHON13708, are clearly detected. For comparison, the fourth panel in the set shows a region around position 11,778 that is identical in both samples.

provide the foundation for a powerful genetic analysis technology. The method can be used to characterize the spectrum of sequence variation in a population and can be applied to the analysis of many genes in parallel. In the case of human mtDNA, we simultaneously analyzed the control region, 13 protein coding genes, 22 tRNA genes, and 2 ribosomal RNA genes. The methods described here can be applied to other research areas in molecular genetics; for example, the ability to identify and sequence polymorphisms provides a basis for genetic mapping. The specificity of oligonucleotide hybridization and the scalability of the method suggests the possibility of a dedicated array that could be used to generate a high-resolution genetic map of an entire genome in a single experiment. Likewise, the concepts and techniques described here have been used to develop approaches for mRNA identification and the large-scale, parallel measurement of expression levels (*24*). Thus, the sequence of a gene, its spectrum of change in the population, its chromosomal location, and its dynamics of expression (all essential to a full understanding of function) can be determined with high-density probe arrays. The challenge now is to synthesize and read probe arrays at even higher density. For example, a 2 cm by 2 cm array, synthesized with probes occupying 1-μm synthesis sites in a 4L tiling, could query the entire coding content of the human genome, estimated at 100,000 genes.

REFERENCES AND NOTES

1. F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
2. A. M. Maxam and W. Gilbert, *ibid.*, p. 560.
3. J. D. Watson and F. H. C. Crick, *Nature* **171**, 737 (1953).
4. W. Bains and G. C. Smith, *J. Theor. Biol.* **135**, 303 (1988); Y. P. Lysov *et al.*, *Dokl. Akad. Nank. SSSR* **303**, 1508 (1988); R. Drmanac, I. Labat, I. Brukner, R. Crkvenjakov, *Genomics* **4**, 114 (1989); E. Southern, U. Maskos, R. Elder, *ibid.* **13**, 1008 (1992); see also R. B. Wallace *et al.*, *Nucleic Acids Res.* **6**, 3543 (1979).
5. S. P. A. Fodor *et al.*, *Science* **251**, 767 (1991).
6. A. C. Pease *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 5022 (1994).
7. In the present format, we can routinely achieve a density of 409,600 synthesis sites in a 1.28 cm by 1.28 cm array. Each 20 μm by 20 μm site contains ~4 × 10^9 functional copies of a specific probe, which corresponds to a mean distance of about 100 Å between probes (M. O. Trulson, D. Stern, R. P. Rava, unpublished results).
8. S. Anderson *et al.*, *Nature* **290**, 457 (1981).
9. The control region of mtDNA is characterized by high amounts of sequence polymorphism concentrated in two hypervariable regions [B. D. Greenberg, J. E. Newbold, A. Sugino, *Gene* **21**, 33 (1983); C. F. Aquardo and B. D. Greenberg, *Genetics* **103**, 287 (1983)].
10. R. L. Cann, W. M. Brown, A. C. Wilson, *Genetics* **106**, 479 (1984).
11. The mt1 and mt2 sequences were cloned from amplified genomic DNA extracted from hair roots [P. Gill, A. J. Jeffreys, D. J. Werrett, *Nature* **318**, 577 (1985); R. K. Saiki *et al.*, *Science* **239**, 487 (1988)]. The clones were sequenced conventionally (*1*). Cloning was performed only to provide a set of pure reference samples of known sequence. For templates for fluorescent labeling, DNA was reamplified from the clones with primers bearing bacteriophage T3 and T7 RNA polymerase promoter sequences (bold; mtDNA sequences uppercase): L15935-T3, 5′-ctcg**gaattaaccctcactaaagg**AAACCTTTTTCCAAGGA and H667-T7, 5′-**taatacgactcactatagggaga**GGCTAGGACCAAACCTATT.
12. Labeled RNAs from the two complementary mtDNA strands [designated L and H (*8*)] were transcribed in separate reactions from a promoter-tagged polymerase chain reaction (PCR) product. Each 10-μl reaction contained 1.5 mM each of the triphosphate nucleotides ATP, CTP, GTP, and UTP; 0.24 mM fluorescein-12–CTP (Du Pont); 0.24 mM fluorescein-12–UTP (Boehringer Mannheim); ~1 to 5 nM (1.5 μl) crude unpurified 1.3-kb PCR product; and T3 or T7 RNA polymerase (1 U/μl) (Promega) in a reaction buffer supplied with the enzyme. The reaction was carried out at 37°C for 1 to 2 hours. RNA was fragmented to an average size of <100 nucleotides by adjusting the solution to 30 mM $MgCl_2$, by the addition of 1 M $MgCl_2$, and heating at 94°C for 40 min. Fragmentation improved the uniformity and specificity of hybridization (M. Chee *et al.*, data not shown). The extent of fragmentation is dependent on the magnesium ion concentration [J. W. Huff, K. S. Sastry, M. P. Gordon, W. E. C. Wacker, *Biochemistry* **3**, 501 (1964); J. J. Butzow and G. L. Eichorn, *Biopolymers* **3**, 95 (1965)]. Good hybridization results have been obtained with both DNA and RNA targets prepared with a variety of labeling schemes, including incorporation of fluorescent and biotinylated deoxynucleoside triphosphates by DNA polymerases, incorporation of dye-labeled primers during PCR, ligation of labeled oligonucleotides to fragmented RNA, and direct labeling by photo-cross-linking a psoralen derivative of biotin directly to fragmented nucleic acids (L. Wodicka, personal communication).
13. For two-color detection experiments, the reference and unknown samples were labeled with biotin and fluorescein, respectively, in separate transcription reactions. Reactions were carried out as described (*12*) except that each contained 1.25 mM of ATP, CTP, GTP, and UTP and 0.5 mM fluorescein-12-UTP or 0.25 mM biotin-16–UTP (Boehringer Mannheim). The two reactions were mixed in the ratio 1:5 (v/v) biotin:fluorescein and fragmented (*12*). Targets were diluted to a final concentration of ~100 to 1000 pM in 3M TMACl [W. B. Melchior Jr. and P. H. von Hippel, *Proc. Natl. Acad. Sci. U.S.A.* **70**, 298 (1973)], 10 mM tris-HCl, pH 8.0, 1 mM EDTA, 0.005% Triton X-100, and 0.2 nM control oligonucleotide labeled at the 5′ end with fluorescein (5′-CTGAACGGTAGCATCTTGAC). Samples were denatured at 95°C for 5 min, chilled on ice for 5 min, and equilibrated to 37°C. A volume of 180 μl of hybridization solution was then added to the flow cell [R. Lipshutz *et al.*, *Biotechniques* **19**, 442 (1995)] and the chip incubated at 37°C for 3 hours with rotation at 60 rpm. The chip was washed six times at room temperature with 6× SSPE (0.9 M NaCl, 60 mM NaH_2PO_4, 6 mM EDTA, pH 7.4), 0.005% Triton X-100. Phycoerythrin-conjugated streptavidin (2 μg/ml in 6× SSPE, 0.005% Triton X-100) was added and incubation continued at room temperature for 5 min. The chip was washed again

Light-Directed, Spatially Addressable Parallel Chemical Synthesis

STEPHEN P. A. FODOR,* J. LEIGHTON READ, MICHAEL C. PIRRUNG,† LUBERT STRYER,‡ AMY TSAI LU, DENNIS SOLAS

See next page

Solid-phase chemistry, photolabile protecting groups, and photolithography have been combined to achieve light-directed, spatially addressable parallel chemical synthesis to yield a highly diverse set of chemical products. Binary masking, one of many possible combinatorial synthesis strategies, yields 2^n compounds in n chemical steps. An array of 1024 peptides was synthesized in ten steps, and its interaction with a monoclonal antibody was assayed by epifluorescence microscopy. High-density arrays formed by light-directed synthesis are potentially rich sources of chemical diversity for discovering new ligands that bind to biological receptors and for elucidating principles governing molecular interactions. The generality of this approach is illustrated by the light-directed synthesis of a dinucleotide. Spatially directed synthesis of complex compounds could also be used for microfabrication of devices.

THE REVOLUTION IN MICROELECTRONICS HAS BEEN MADE possible by photolithography, a process in which light is used to spatially direct the simultaneous formation of many electrical circuits. We report a method that uses light to direct the simultaneous synthesis of many different chemical compounds. Synthesis occurs on a solid support. The pattern of exposure to light or other forms of energy through a mask, or by other spatially addressable means, determines which regions of the support are activated for chemical coupling. Activation by light results from the removal of photolabile protecting groups from selected areas (Fig. 1). After deprotection, the first of a set of "building blocks" (for example, amino acids or nucleic acids, each bearing a photolabile protecting group) is exposed to the entire surface, but reaction occurs only with regions that were addressed by light in the preceding step. The substrate is then illuminated through a second mask, which activates a different region for reaction with a second protected building block. The pattern of masks used in these illuminations and the sequence of reactants define the ultimate products and their locations. The number of compounds that can be synthesized by this technique is limited only by the number of synthesis sites that can be addressed with appropriate resolution. Combinatorial masking strategies can be used to form a large number of compounds in a small number of chemical steps. Moreover, a high degree of miniaturization is possible because the density of synthesis sites is bounded only by physical limitations on spatial addressability, in this case the diffraction of light. Each compound is accessible and its position is precisely known. Hence, its interactions with other molecules can be assessed.

Spatially localized photodeprotection. Spatially localized substrate activation can be accomplished by photolithographic techniques. Amino groups at the ends of linkers attached to a glass substrate were derivatized with nitroveratryloxycarbonyl (NVOC), a photoremovable protecting group (*1*). Photodeprotection was effected by illumination of the substrate through a mask (a 100 μm by 100 μm checkerboard) with alternating opaque and transparent elements. The free amino groups were fluorescently labeled by treatment of the entire substrate surface with fluorescein isothiocyanate (FITC). The substrate was then scanned in an epifluorescence microscope. The presence of a high-contrast fluorescent checkerboard pattern with 100 μm by 100 μm elements (depicted in red in Fig. 2) reveals that free amino groups were generated in specific regions by spatially localized photodeprotection.

Light-directed peptide synthesis. Light-directed synthesis of two pentapeptides was carried out as outlined in Fig. 3. The 1-hydroxybenzotriazole (HOBt)–activated ester of NVOC-Leu (NVOC-Leu-OBt) was allowed to react with the entire surface of a substrate that had previously been derivatized with amino functional groups. After removal of the NVOC protecting group by uniform illumination, the substrate was treated with NVOC-Phe-OBt. Two repetitions of this cycle with NVOC-Gly-OBt generated a substrate containing NVOC-GGFL across the entire surface (*2*). Spatially localized photodeprotection was then performed through a 50-μm checkerboard mask. The surface was then treated with *Nα-tert*-butyloxy carbonyl-*O-tert*-butyl-L-tyrosine. Finally, the surface was uniformly illuminated to photolyze the remaining NVOC-GGFL sites and treated with NVOC-Pro-OBt. After removal of the protecting groups, the surface consists of an array of H_2N-Tyr-Gly-Gly-Phe-Leu (YGGFL) and H_2N-Pro-Gly-Gly-Phe-Leu (PGGFL) peptides in 50 μm by 50 μm elements.

Antibody recognition of the peptide pattern. The pentapeptide array was probed with a mouse monoclonal antibody directed against β-endorphin. This antibody (called 3E7) binds YGGFL and YGGFM with nanomolar affinity (*3*) and requires the amino-terminal Tyr for high-affinity binding. A second incubation with fluorescein-labeled goat antibody to mouse was used to detect regions containing bound 3E7. As shown in Fig. 4, a high-contrast (>12:1 intensity ratio) fluorescence checkerboard image shows that

The authors are at the Affymax Research Institute, 3180 Porter Drive, Palo Alto, CA 94304.

*To whom correspondence should be addressed.
†Present address: Department of Chemistry, Duke University, Durham, NC 27706.
‡Present address: Department of Cell Biology, Stanford University School of Medicine, Stanford, CA 94305.

group of deoxycytidine (see Fig. 8).

The 3-D representation of the fluorescence intensity data in Fig. 8 reproduces the checkerboard illumination pattern used during photolysis of the substrate. This result demonstrates that oligonucleotides as well as peptides can be synthesized by the light-directed method.

Comparison to other methods and potential applications. We have introduced an approach for the simultaneous synthesis of a large number of compounds that combines solid-phase synthesis (*11*), photolabile protecting groups (*1*), and photolithography (*12*). The method can be applied to any solid-phase synthesis technique in which light can be used to generate a reactive group. We have used light-directed spatially addressable parallel chemical synthesis to synthesize large arrays of peptides. The light-directed formation of oligonucleotides attests to the versatility of the technique and suggests that it could be broadly applicable in making high-density arrays of chemical compounds. The ten-step binary synthesis results in the formation of 1024 peptides in 1.6 cm^2. The 50-μm checkerboard pattern of alternating pentapeptides shows that 40,000 compounds can be synthesized in 1 cm^2. Our present capability for high-contrast photodeprotection is better than 20 μm, which gives >250,000 synthesis sites per square centimeter. There is no physical reason why higher densities of synthesis sites cannot be achieved. Indeed, high spatial resolution electron-beam lithography (~250 Å) has been used to generate patterns at a density of 10^{10} per square centimeter (*13*).

It is interesting to compare the light-directed method with other techniques for parallel chemical synthesis. One approach is to physically segregate different reactants by pipeting them into different reaction vessels. For example, 96 peptides have been simultaneously synthesized on the tips of pins by immersing them into different solutions that are contained in the chambers of a microtiter plate (*14*). The need for physical separation of reaction sites sharply limits the number of compounds that can be made by the pin method. In contrast, very large numbers of peptides can be generated by recombinant DNA approaches (*9, 15*). Millions of different peptide sequences can be expressed on the surface of phage by inserting randomly synthesized oligonucleotides into their genomes. Each phage clone displays a different peptide. Although the peptides-on-phage are in suspension and are not fixed at defined locations, those that bind tightly to a receptor can be identified by panning, isolation of individual clones, and DNA sequencing. Only peptides that contain genetically coded amino acids can be generated by expression on phage. The recombinant and light-directed approaches have distinctive strengths that are complementary. For example, a peptide identified by the phage method to have appreciable affinity for a receptor can serve as the kernel around which diversity is generated by light-directed synthesis. A synthesis might include custom chemical building blocks in addition to the standard set of L-amino acids (*16*). For example, D-amino acids could be introduced to make the peptide more resistant to proteolysis (*17*), and modified site chains (*18*) could be inserted to increase affinity.

Parallel chemical synthesis could be used to explore molecular recognition processes in biology and other fields. For example, pharmaceutical discovery is increasingly based on an understanding of the way receptors and enzymes interact with specific ligands. The techniques described here allow the synthesis of large numbers of peptides or other oligomers that can be surveyed for binding to biological macromolecules.

Fabrication of small devices such as microelectronic circuits relies on the chemistry of photoresists, vapor deposition, and ion implantation. The techniques described here enable the in situ synthesis of complex compounds on a microscale. The methods of spatially addressable chemical synthesis may be used in conjunction with the microfabrication of circuitry. The union of these technologies may find applications in novel detection devices containing arrays of biological receptors or other molecular recognition elements.

The functional properties of molecules synthesized by the light-directed approach can be read in a variety of ways. As was shown here, the binding of a receptor such as an antibody can readily be detected fluorimetrically. Radioactive or chemiluminescent labels could also be used (*19*). The susceptibility of compounds in an array to modification by an enzyme or other catalyst could also be directly assayed. For example, the cleavage of a peptide at a site located between a fluorescent energy donor and acceptor would lead to increased fluorescence (*20*). Peptides that are effective substrates for phosphorylation by a kinase could be identified by monitoring the ^{32}P pattern following incubation with enzyme and radiolabeled ATP (adenosine triphosphate).

Oligonucleotide arrays produced by light-directed synthesis could be used to detect complementary sequences in DNA and RNA. Such arrays would be valuable in gene mapping, fingerprinting, diagnostics, and nucleic acid sequencing. A sequencing method based on hybridization to a complete set of fixed-length oligonucleotides immobilized individually as dots of a two-dimensional matrix has been proposed (*21*). It is noteworthy that the light-directed synthesis of all 65,536 possible octanucleotides (4^8) would fit into 1.6 cm^2 with 50-μm square sites, a resolution already achieved.

REFERENCES AND NOTES

1. A. Patchornik, B. Amit, R. B. Woodward, *J. Am. Chem. Soc.* 92, 6333 (1970). There is a wide array of available photochemical protecting groups [V. N. R. Pillai, *Synthesis* 1980, 1 (1980)]. We have used NVOC because of its favorable absorption characteristics and its established use in peptide synthesis.
2. One-letter codes for the amino acids (L-isomers except for G, Gly): A, Ala; F, Phe; L, Leu; M, Met; P, Pro; Q, Gln; S, Ser; T, Thr; and Y, Tyr. Lower-case one-letter codes are reserved for the D-isomers.
3. T. Meo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 80, 4084 (1983).
4. Compounds formed in a light-activated synthesis can be positioned in any defined geometric array as long as equivalent transformations are used for each row. A square or rectangular matrix is convenient but not required.
5. Not all light-activated syntheses can be represented as factored polynomials. Some can only be denoted by irreducible (prime) polynomials.
6. Binary rounds and nonbinary rounds can be interspersed as desired, as in

$$P = (A + \varnothing)(B)(C + D + \varnothing)(E + F + G)$$

The 18 compounds formed are ABCE, ABCF, ABCG, ABDE, ABDF, ABDG, ABE, ABF, ABG, BCE, BCF, BCG, BDE, BDF, BDG, BE, BF, and BG. The switch matrix S for this seven-step synthesis is

$$S = \begin{matrix} 1111111111000000000 \\ 1111111111111111111 \\ 1111000000111000000 \\ 0001110000001111000 \\ 1001001001001001001 \\ 0100100100100100100 \\ 0010010010010010010 \end{matrix}$$

The round denoted by (B) places B in all products because the reaction area was uniformly activated (the mask for B consisted entirely of 1's). The number of compounds k formed in a synthesis consisting of r rounds, in which the ith round has b_i chemical reactants and z_i nulls, is

$$k = \prod (b_i + z_i),$$

and the number of chemical steps n is

$$n = \Sigma\, b_i.$$

The number of compounds synthesized when $b = a$ and $z = 0$ in all rounds is $a^{n/a}$, compared with 2^n for a binary synthesis. For $n = 20$ and $a = 5$, 625 compounds (all tetramers) would be formed, compared with 1.049×10^6 compounds in a binary synthesis with the same number of chemical steps. It should also be noted that rounds in a polynomial can be nested, as in

$$[(A + (B + \varnothing)(C + \varnothing)](D + \varnothing)$$

The products are AD, BCD, BD, CD, D, A, BC, B, C, and $\varnothing$.
7. The longest peptide formed is fYGAGTFLSF (the amino-terminal is f and the carboxyl-terminal residue linked to the substrate is F). Because the solid-phase synthesis is carried out in the carboxyl-to-amino direction, F is the first, S is the second, and f is the tenth unit to be coupled.
8. In cases where the monovalent interaction of a receptor with a ligand is very low (where the off rate is rapid), it may not be possible to detect the binding of the

Concise review paper / *Mise au point*

Human genome diversity

Diversité génomique humaine

Howard M. Cann

Fondation Jean-Dausset, Centre d'étude du polymorphisme humain (CEPH), 27, rue Juliette-Dodu, 75010 Paris, France

(Received 20 March 1998, accepted after revision 23 March 1998)

Abstract – Human genome diversity studies analyse genetic variation among individuals and between populations in order to understand the origins and evolution of anatomically modern humans (*Homo sapiens sapiens*). The availability of thousands of DNA polymorphisms (genetic markers) brings analytic power to these studies. Human genome diversity studies have clearly shown that the large part of genetic variability is due to differences among individuals within populations rather than to differences between populations, effectively discrediting a genetic basis of the concept of 'race'. Evidence from paleontology, archaeology and genetic diversity studies is quite consistent with an African origin of modern humans more than 100 000 years ago. The evidence favors migrations out of African as the source of the original peopling of Asia, Australia, Europe and Oceania. An international program for the scientific analysis of human genome diversity and of human evolution has been developed. The Human Genome Diversity Project (HGDP) aims to collect and preserve biologic samples from hundreds of populations throughout the world, make DNA from these samples available to scientists and distribute to the scientific community the results of DNA typing with hundreds of genetic markers. (© Académie des sciences / Elsevier, Paris.)

human genome diversity / genetic variability / *Homo sapiens sapiens* / modern human origins and evolution / polymorphic DNA markers / Human Genome Diversity Project.

Résumé – Les études de la diversité génomique humaine analysent la variation génétique entre individus et entre populations afin de comprendre l'évolution de l'être humain moderne (*Homo sapiens sapiens*). La mise en évidence d'un grand nombre de polymorphismes de l'ADN nucléaire et mitochondrial a permis l'affinement de ces analyses. Ces études ont montré que la majeure partie de la variabilité génétique était due aux différences entre individus d'une même population plutôt qu'entre les populations elles-mêmes, infirmant ainsi une base génétique du concept de « race ». Les études de paléontologie, d'archéologie et de diversité génétique ont montré de façon cohérente une origine africaine des êtres humains modernes et ceci, il y a plus de 100 000 ans. Certains d'entre eux ont ensuite migré vers l'Asie, l'Australie, l'Europe et l'Océanie. Un programme international ayant pour objectif l'analyse scientifique de la diversité génomique et de l'évolution humaine est actuellement mis en place. Ce programme appelé Diversité génomique humaine a pour but de recueillir et de conserver les échantillons biologiques de centaines de populations dans le monde entier, de fournir aux chercheurs de l'ADN de ces échantillons pour les caractériser avec des centaines de marqueurs génétiques, et d'analyser ces données génétiques qui seront ensuite mises à la disposition de la communauté scientifique. (© Académie des sciences / Elsevier, Paris.)

diversité génomique humaine / variabilité génétique / *Homo sapiens sapiens* / origine et évolution de l'être humain moderne / marqueurs génétiques de l'ADN

Note communicated by Jean Rosa

E-mail: howard.cann@cephb.fr

shown that indeed language and allele frequencies tend to be correlated among populations. Intriguingly, there are now hints in several unilingual populations of differing Y chromosome and mtDNA contributions to genetic variation, in that the variation of the latter marker is increased as compared to that of the former. The conclusion here may be that the linguistic barrier is more readily breached by women through incorporation into a predominantly unilingual population.

Probably the most recent dramatic and novel result in human genetic diversity studies has been the sequencing of a portion of the control region of mtDNA extracted with great care from a *Homo sapiens neanderthalensis* fossil thought to be between 30 000 and 100 000 years old. This incredible experiment, following an admirably rigorous design, provided the entire 379-bp sequence of the hypervariable region 1, as deduced from many cloned, overlapping, short PCR products. This sequence was shown to differ markedly from the corresponding sequence of all known modern human mtDNAs. The mean number of pairwise base substitution differences between the Neanderthal and modern human mtDNA hypervariable region 1 (27 differences) is more than three times that observed among humans (on average eight differences). The difference seems to be sufficient to place the Neanderthal sequence outside of the variation that occurs among humans. This is an exciting result indicating that Neanderthals did not contribute mtDNA (and presumably nuclear DNA) to modern humans.

The above examples of results issuing from studies of human genetic/genome diversity show the importance in human biology of this field of research. The discrediting of a genetic basis of the concept of race, understanding the origin of modern humans and the details of the peopling of the world and the sequencing of Neanderthal mtDNA are hardly trivial undertakings, and there are many other interesting and important questions to be posed and answered. For instance, haplotypes, which are more informative than individual loci for the description of chromosomes of population founders, will become the genetic units for analyses of human genome diversity, which, in turn, will provide information on their origins, ages and evolution. The development of molecular polymorphic markers, and many of them, provides a depth of analytic resolution and power heretofore unavailable for, and is clearly impinging on, research design of diversity studies. The concept of genome diversity is clearly embodied in developing future studies involving markers drawn from throughout the genome, hundreds of markers that are highly polymorphic, as well as thousands of the more stable (less mutation and thus less polymorphic) single nucleotide polymorphisms (SNPs) that detect variation on-average once every ~1 000 base pairs. Automatic typing of both groups of markers is reality and allows the equivalent of diversity genome scans of thousands of individuals. Put these together with methods for high throughput automatic DNA extraction from thousands of blood samples collected from world-wide population samples of hundreds of individuals each and with analytic methods for calculating inter-population genetic distances and evolutionary trees and describing in detail geographic variation of populations, essentially based on functions of allele frequency distributions, and one has the ingredients for an organized international collaboration on human genome diversity. Indeed, the Human Genome Diversity Project, after a slow start, is gathering steam, impelled recently by a favorable evaluation of the field by a committee of outstanding scientists and ethics specialists convened by the U.S. National Research Council.

The Human Genome Diversity Project (HGDP) is a program for the scientific analysis of human genetic diversity and evolution. It aims to 1) collect and preserve biologic samples from populations throughout the world; 2) make DNA from these samples available to scientists; and 3) distribute to the scientific community the DNA typing results. The HGDP will be organized as an international collaboration of scientists who work on human variation (usually geneticists, physical anthropologists, paleontologists and archaeologists). Collaborators will provide blood samples from world populations and/or type the DNA from these samples. Collaborator activities will be coordinated by several major international repositories, which will be responsible for receiving and processing blood samples, storing purified lymphocytes and the leukocyte fraction from peripheral blood, establishing lymphoblastoid cell lines (LCLs) and extracting DNA from these resources for distribution to collaborators. A database, containing DNA typing results as well as ethnographic information, will be developed and maintained online for collaborating scientists initially and eventually for the public.

Ethical issues play a critical role in the research design and organization of this project. Protection of the autonomy, privacy and welfare of those who participate in the project has been a central concern of those involved in this type of research. These obligations as they apply to individual subjects and, perhaps to populations, have been discussed and studied by the organizers of the proposed project, as well as by a subcommittee of the UNESCO International Bioethics Committee and the U.S. National Research Council. The project requires a challenging application of the ethical principles used in other aspects of human genetics research.

A preliminary project is planned that would bring together some 500–1 000 already-existing LCLs from populations in Africa, Europe, Asia, the Americas and Oceania. DNA from these LCLs will be distributed to collaborating scientists for testing with various microsatellite and, perhaps, SNP markers in order to develop a common panel of hundreds (to thousands) of markers for use in the HGDP program. These cell lines are expected to be gathered this year. The research program will then follow. The goals of the extended research program are to obtain blood samples from 100 to 250 individuals from

env sequences in each specimen were examined by phylogenetic analysis.

11. The methods described by G. H. Learn *et al.* [*J. Virol.* **70**, 5720 (1996)] were used to align DNA sequences (with the use of CLUSTALW plus manual adjustment), calculate genetic distances (with the use of DNADIST, using the maximum likelihood method), evaluate potential sample mixups, construct neighbor joining trees, and perform bootstrap analyses (1000 replicates). Sequence regions that could not be unambiguously aligned were removed from subsequent analyses. Each sequence was compared for phylogenetic relatedness to the entire set of published and available unpublished laboratory HIV database sequences. If after this analysis the viral sequences from a mother and an infant appeared as a monophyletic group on a phylogenetic tree, they were judged to be phylogenetically linked or to have a common ancestor not shared by sequences from any other individuals evaluated. Issues regarding the assignment of phylogenetic linkage are discussed in greater detail by Learn *et al.*
12. L. M. Frenkel *et al.*, at www.sciencemag.org/feature/data/974996.shl.
13. R. Liu *et al.*, *Cell* **86**, 367 (1996).
14. L. M. Frenkel *et al.*, unpublished data.
15. M.-L. Newell *et al.*, *Lancet* **347**, 213 (1996).
16. P. Palumbo, J. Skurnick, D. Lewis, M. Eisenberg, *J. Acquir. Immune Defic. Syndr. Hum. Retrovirol.* **10**, 436 (1995).
17. A. McMichael, R. Koup, A. J. Ammann, *N. Eng. J. Med.* **334**, 801 (1996).
18. E. C. Holmes *et al.*, *J. Infect. Dis.* **167**, 1411 (1993).
19. T. Liu *et al.*, *J. Immunol.* **154**, 3147 (1995).
20. A. Hoffenbach *et al.*, *ibid.* **142**, 452 (1989).
21. G. Schochetman, S. Subbarao, M. L. Kalish, in *Viral Genome Methods*, K. W. Adolph, Ed. (CRC Press, Boca Raton, FL, 1996), pp. 25–41.
22. E. L. Delwart, M. P. Busch, M. L. Kalish, J. W. Mosley, J. I. Mullins, *AIDS Res. Hum. Retrovir.* **11**, 1181 (1995).
23. C. H. Contag *et al.*, *J. Virol.* **71**, 1292 (1997).

24. We thank J. Conroy for performing PCR assays; E. Abrams, M. S. Orloff, R. C. Reichman, L. M. Demeter, J. S. Lambert, R. Dolin, R. Sperling, D. Shapiro, G. McSherry, and the Ariel Project and ACTG 076 investigators for critical patient specimens; D. Swofford for use of computer program PAUP*, version 4.0.0d63; and C. B. Wilson and K. K. Holmes for editorial contributions. This work was supported by grants from the Pediatric AIDS Foundation (500153-10-PGT, 50366-14-PGR, 55516-ARI, 55529-ARI, 55525-ARI, 55532-ARI, 55526-ARI, 55531-ARI, and 55522-ARI), the U.S. Public Health Service (UO1-27658, AI32910, AI27757, and AI35539), and the Foster Foundation.

22 December 1997; accepted 26 March 1998

Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome

David G. Wang, Jian-Bing Fan, Chia-Jen Siao, Anthony Berno, Peter Young, Ron Sapolsky, Ghassan Ghandour, Nancy Perkins, Ellen Winchester, Jessica Spencer, Leonid Kruglyak, Lincoln Stein, Linda Hsie, Thodoros Topaloglou, Earl Hubbell, Elizabeth Robinson, Michael Mittmann, Macdonald S. Morris, Naiping Shen, Dan Kilburn, John Rioux, Chad Nusbaum, Steve Rozen, Thomas J. Hudson, Robert Lipshutz,* Mark Chee, Eric S. Lander*

Single-nucleotide polymorphisms (SNPs) are the most frequent type of variation in the human genome, and they provide powerful tools for a variety of medical genetic studies. In a large-scale survey for SNPs, 2.3 megabases of human genomic DNA was examined by a combination of gel-based sequencing and high-density variation-detection DNA chips. A total of 3241 candidate SNPs were identified. A genetic map was constructed showing the location of 2227 of these SNPs. Prototype genotyping chips were developed that allow simultaneous genotyping of 500 SNPs. The results provide a characterization of human diversity at the nucleotide level and demonstrate the feasibility of large-scale identification of human SNPs.

Although the Human Genome Project still has tremendous work ahead to produce the first complete reference sequence of the human chromosomes, attention is already focusing on the challenge of large-scale characterization of the sequence variation among individuals (*1*). This genetic diversity is of interest because it explains the basis of heritable variation in disease susceptibility, as well as harbors a record of human migrations.

The most common type of human genetic variation is the SNP, a position at which two alternative bases occur at appreciable frequency (>1%) in the human population. There has been growing recognition that large collections of mapped SNPs would provide a powerful tool for human genetic studies (*1*, *2*). SNPs can serve as genetic markers for identifying disease genes by linkage studies in families, linkage disequilibrium in isolated populations, association analysis of patients and controls, and loss-of-heterozygosity studies in tumors (*1*, *2*). Although individual SNPs are less informative than currently used genetic markers (*3*), they are more abundant and have greater potential for automation (*4*, *5*).

We performed an initial survey to identify SNPs by using conventional gel-based DNA sequencing to examine sequence-tagged sites (STSs) distributed across the human genome. STSs are short genomic sequences that can be amplified from DNA samples by means of a corresponding polymerase chain reaction (PCR) assay. From among 24,568 STSs used in the construction of a physical map of the human genome at the Whitehead Institute for Biomedical Research/MIT Center for Genome Research (*6*, *7*), an initial collection of 1139 STSs was chosen (*8*). These STSs contained a total of 279 kb of genomic sequence (*9*), with one-third from random genomic sequence and two-thirds from 3′-ends of expressed sequence tags (3′-ESTs) and primarily representing untranslated regions of genes. Each STS was amplified from four samples (*10*): three individual samples and a pool of 10 individuals (thereby permitting allele frequencies to be estimated among 20 chromosomes). The PCR products were subjected to single-pass DNA sequencing based on fluorescent-dye primers and gel electrophoresis; sequence traces were compared by a computer program followed by visual inspection (*11*). Candidate SNPs were declared when two alleles were seen among the three individuals, with both alleles present at a frequency greater than 30% in the pooled sample. The term "candidate SNP" is used because a subset of such apparent polymorphisms turn out to be sequencing artifacts, as discussed below.

The survey identified 279 candidate SNPs, distributed across 239 of the STSs. This corresponds to a rate of one SNP per 1001 base pairs (bp) screened and an observed nucleotide heterozygosity of $H = 3.96 \times 10^{-4}$ (Table 1). Expressed sequences (3′-ESTs) showed a lower polymorphism rate than random genomic sequence (with

D. G. Wang, C.-J. Siao, P. Young, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, E. Robinson, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA.
J.-B. Fan, A. Berno, R. Sapolsky, G. Ghandour, L. Hsie, T. Topaloglou, E. Hubbell, M. Mittmann, M. S. Morris, N. Shen, R. Lipshutz, M. Chee, Affymetrix, Incorporated, 3380 Central Expressway, Santa Clara, CA 95051, USA.
E. S. Lander, Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA, and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
*To whom correspondence should be addressed.

seen in a test set of 39 individuals fell into distinct clusters, corresponding to the possible genotypes (28). These clusters could then be used to assign genotypes for further samples (29).

The cluster test was applied to the ~500 candidate SNPs that worked well under multiplex amplification conditions: 75% passed the cluster test, and careful resequencing demonstrated that all such loci were true polymorphisms. The cluster test thus provides reliable confirmation of an SNP. The remaining 25% failed the cluster test, and resequencing revealed that half were false positives in the SNP screen and half were true polymorphisms (with the poor discrimination on the chip typically due to one allele hybridizing more weakly than the other). Thus, 88% of the candidate SNPs proved to be true polymorphisms, and 86% of true SNPs passed the cluster test.

To test the reproducibility and accuracy of the genotyping method, we genotyped a set of 91 loci (passing the cluster test) in three individuals by performing chip-based genotyping on six separate occasions over a 2-month period. The correct genotypes were independently determined by thorough gel-based resequencing. The genotyping-chip assay assigned a genotype in 98% of cases (1613/1638), and this assignment proved correct in 99.9% (1611/1613) of these cases. The loci were also genotyped in two complete CEPH families. The genotypes were not independently confirmed, but they were fully consistent with mendelian segregation.

For SNPs passing the cluster test, highly accurate genotypes could thus be obtained with the simple design used here. For the remaining SNPs (14%), similar accuracy can likely be obtained but may require optimization of the genotyping array design, depending on the locus [as shown in (5)].

The SNP surveys provide data about human genetic diversity. Two classical measures of diversity (30) are H, the average heterozygosity per nucleotide, and K, the proportion of sites harboring a variation. H does not depend on sample size, whereas K increases with the number of genomes surveyed. For a population at equilibrium, the neutral theory of evolution relates H and K to the classical population genetic parameter $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate per nucleotide. (θ can be thought of as twice the number of new mutations per generation arising in a population with size N_e.) Specifically, $H \approx \theta$ and $K \approx \theta [1^{-1} + 2^{-1} + 3^{-1} + \ldots + (n-1)^{-1}]$, provided that θ is small. From these equations, one can estimate θ based on H or K.

The human population is not at equilibrium, but rather underwent a rapid population expansion in the last 100,000 to 200,000 years. Such population explosions tend to suppress the effects of genetic drift and thus preserve the distribution of common alleles and the value of θ. Accordingly, the value of θ is relevant to the ancestral human population before its recent expansion.

The four estimates of θ derived from H and K for the two surveys are all roughly $\theta \approx 4 \times 10^{-4}$ (Table 1). Assuming a mutation frequency of $\mu \approx 10^{-8}$ to 10^{-9}, this would suggest an effective population size of $N_e \approx 10^4$ to 10^5, which seems reasonable for the ancestral population preceding the explosion in the last 100,000 years (31). Strictly speaking, these estimates apply only to the European population, from which all samples were drawn. However, a preliminary survey of a more diverse sample of 31 individuals representing all major racial groups yielded a value of θ that is only 30% larger (26), consistent with the idea that human variation occurs primarily within rather than between racial groups (32).

The resources reported here represent only a first step toward a dense SNP map of the human genome. The genetic map should already be useful for family-based linkage studies, given the average spacing (2 cM) and average heterozygosity (34%) of the markers. (The heterozygosity applies to the European-derived samples studied here, but a preliminary survey of ~180 of the SNPs shows that most are also polymorphic in other groups.) It still remains to develop a suitable genotyping system, such as a 2000-SNP genotyping chip.

Large-scale screening for human variation is clearly feasible. Someday it may become possible to screen entire human genomes. In the nearer term, a key goal will be to extend SNP discovery to the protein coding regions of all human genes (roughly 120 Mb of sequence, only about 40 times more than the current study) in order to catalog the common variants that may explain susceptibility to common genetic traits and diseases (1).

REFERENCES AND NOTES

1. N. Risch and K. Merikangas, *Science* **273**, 1516 (1996); E. S. Lander, *ibid.* **274**, 536 (1996); F. S. Collins, M. S. Guyer, A. Chakravarti, *ibid.* **278**, 1580 (1997).
2. L. Kruglyak, *Nature Genet.* **17**, 21 (1997).
3. SNPs have only two alleles and are less informative than typical multi-allelic simple sequence length polymorphisms (SSLPs). This disadvantage can be offset by using a greater density of SNPs: a genome scan with 1000 well-spaced SNPs, for example, will extract about the same linkage information as the current standard of 400 well-spaced SSLPs (2).
4. B. J. Conner *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 278 (1983); U. Landegren, R. Kaiser, J. Sanders, L. Hood, *Science* **241**, 1077 (1988); D. Y. Wu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2757 (1989); R. K. Saiki *et al.*, *ibid.*, p. 6230; A.-C. Syvanen *et al.*, *Genomics* **8**, 684 (1990); D. A. Nickerson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 8923 (1990); K. J. Livak *et al.*, *Nature Genet.* **9**, 341 (1995); M. T. Roskey *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 4724 (1996).
5. M. T. Cronin *et al.*, *Hum. Mutat.* **7**, 244 (1996).
6. T. J. Hudson *et al.*, *Science* **270**, 1945 (1995).
7. G. D. Schuler *et al.*, *ibid.* **274**, 540 (1996).
8. STSs with the largest sizes were used in the gel-based screen, and the remaining STSs, having somewhat smaller sizes, were used in the subsequent chip-based screen.
9. The genomic sequence screened (279 kb) is the sum of the distances between the primer sites of the STSs successfully resequenced.
10. The individuals surveyed were chosen from Centre d'Etude du Polymorphisme Humain (CEPH) pedigrees K104, K884, and K1331 from the Amish, Venezuelan, and Utah populations, respectively. The SNP survey by gel-based sequencing examined three unrelated individuals (K104-1, K884-2, K1331-1) and a pool of 10 individuals (K104-13, -14, -15, -16; K884-15, -16; K1331-12, -13, -14, -15). The SNP survey by chip-based analysis examined seven unrelated individuals (K104-1, -16; K884-2, -15, -16; K1331-12, -13).
11. STSs were amplified with their corresponding PCR primers as described (6), except that the forward primer was modified to include the M13 −21 primer site (5′-TGTAAAACGACGGCCAGT-3′) at its 5′-end. The resulting PCR products were subjected to dye-primer sequencing (33), with products detected on an ABI377 or ABI373 fluorescence sequence detector. Possible sequence variations were detected by the ABI Sequence Navigator software package, which suggests potential heterozygotes by identifying nucleotide positions at which a secondary peak exceeds a selected threshold (50%). Such apparent variations were then visually inspected to compare the patterns seen among the several individuals.
12. D. N. Cooper and M. Karwczak, *Hum. Genet.* **85**, 55 (1990).
13. M. Chee *et al.*, *Science* **274**, 610 (1996); M. J. Kozal *et al.*, *Nature Med.* **2**, 753 (1996).
14. S. P. A. Fodor *et al.*, *Science* **251**, 767 (1991); A.-C. Pease *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 5022 (1994). The current generation of technology allows fabrication of 1.28 cm by 1.28 cm arrays of ~320,000 distinct oligonucleotides, each residing in a "feature" of ~20 μm by 25 μm and containing $>10^7$ copies of the probe.
15. J. G. Hacia *et al.*, *Nature Genet.* **14**, 441 (1996).
16. STSs were amplified with their corresponding PCR primers as described (6). PCR products intended for hybridization to the same chip (typically 100 to 200 STSs from a single individual) were pooled together for subsequent processing. About 1 to 2 μg of the pooled PCR product was purified with Qiaquick purification kit (Qiagen), fragmented with deoxyribonuclease (DNase) I (Promega) and labeled with biotin with terminal deoxynucleotidyl transferase (TdT, GibcoBRL Life Technology). The purification was performed according to the manufacturer's instructions. The fragmentation was performed in a 40-μl reaction with 0.2 unit of DNase I, 10 mM tris-acetate (pH 7.5), 10 mM magnesium acetate, and 50 mM potassium acetate at 37°C for 15 min, after which the reaction was stopped by heat inactivation at 96°C for 15 min. The terminal transferase reaction was performed by adding 15 units of TdT and 12.5 μM biotin-N6-ddATP (DuPont NEN) to the preceding reaction mixture, incubating it at 37°C for 1 hour, and then heat-inactivating it at 96°C for 15 min. The labeled samples were hybridized to the chip as follows. Samples were denatured at ~96°C for 5 to 6 min and cooled on ice for 2 to 5 min. Chips were first hybridized with 6× SSPET [0.9 M NaCl, 60 mM NaH_2PO_4, 6 mM EDTA (pH 7.4), 0.005% Triton X-100] for ~5 min and then hybridized with the denatured sample in hybridization buffer [3M tetramethylammonium chloride, 10 mM tris-HCl (pH 7.8), 1 mM EDTA, 0.01% Triton X-100, herring sperm DNA (100 μg/ml), and 200 pM control oligomer] at 44°C for 15 hours on a rotisserie at 40 rpm. Chips were washed three times with 1× SSPET, 10 times

High density synthetic oligonucleotide arrays

Robert J. Lipshutz, Stephen P.A. Fodor, Thomas R. Gingeras & David J. Lockhart

Affymetrix, Inc. 3380 Central Expressway, Santa Clara, California 95051, USA. e-mail: rob_lipshutz@affymetrix.com

Experimental genomics involves taking advantage of sequence information to investigate and understand the workings of genes, cells and organisms. We have developed an approach in which sequence information is used directly to design high-density, two-dimensional arrays of synthetic oligonucleotides. The GeneChip® probe arrays are made using spatially patterned, light-directed combinatorial chemical synthesis, and contain up to hundreds of thousands of different oligonucleotides on a small glass surface. The arrays have been designed and used for quantitative and highly parallel measurements of gene expression, to discover polymorphic loci and to detect the presence of thousands of alternative alleles. Here, we describe the fabrication of the arrays, their design and some specific applications to high-throughput genetic and cellular analysis.

Biological systems read, store and modify genetic information using the rules of molecular recognition. Every nucleic acid strand carries the capacity to recognize complementary sequences through base pairing. The process of recognition, or hybridization, can be highly parallel; every sequence in a complex mixture can, in principle, be interrogated simultaneously. We have used these simple principles to develop powerful new experimental tools designed to collect and analyse vast amounts of genetic and cellular information. The introduction, development and integration of two key technologies[1–5] form the cornerstone of the new methods. The first is the fabrication of hundreds of thousands of polynucleotides at high spatial resolution in precise locations on a surface. The second, laser confocal fluorescence scanning, facilitates the measurement of molecular binding events on the array. These technologies and some variants have been adopted in both the commercial and academic sectors (see pages 25 (ref. 6), 10 (ref. 7) and 15 (ref. 8) of this issue).

At Affymetrix, we have focused on light-directed synthesis for the construction of high-density DNA probe arrays using two techniques: photolithography and solid-phase DNA synthesis. We attach synthetic linkers modified with photochemically removable protecting groups to a glass substrate and direct light through a photolithographic mask to specific areas on the surface to produce localized photodeprotection (Fig. 1). The first of a series of chemical building blocks, hydroxyl-protected deoxynucleosides, is incubated with the surface, and chemical coupling occurs at those sites that have been illuminated in the preceding step. Next, light is directed to different regions of the substrate by a new mask, and the chemical cycle is repeated[9,10]. Highly efficient strategies can be used to synthesize arbitrary polynucleotides at specified locations on the array in a minimum number of chemical steps[1]. For example, the complete set of 4^N polydeoxynucleotides of length N, or any subset, can be synthesized in only 4×N cycles. Thus, given a reference sequence, a DNA probe array can be designed that consists of a highly dense collection of complementary probes with virtually no constraints on design parameters. The amount of nucleic acid information encoded on the array in the form of different probes is limited only by the physical size of the array and the achievable lithographic resolution. Current large scale commercial manufacturing methods allow for approximately 300,000 polydeoxynucleotides to be synthesized on small 1.28×1.28 cm arrays—experimental versions now exceed one million probes per array.

Photolithography allows the construction of arrays with extremely high information content. Because the arrays are constructed on a rigid material (glass), they can be inverted and mounted in a temperature-controlled hybridization chamber. A fluorescently tagged nucleic acid sample injected into the chamber hybridizes to complementary oligonucleotides on the array. Laser excitation enters through the back of the glass support, focused at the interface of the array surface and the target solution. Fluorescence emission is collected by a lens and passes through a series of optical filters to a sensitive detector. By simply scanning the laser beam or translating the array, or a combination of both, a quantitative two-dimensional fluorescence image of hybridization intensity is quickly obtained[1,2].

Gene expression monitoring

Once sequence information (partial or complete) for a gene is obtained, the next question is generally: "what does its product do?". To understand gene function, it is helpful to know when and where it is expressed, and under what circumstances the expression level is affected. Beyond questions of individual gene

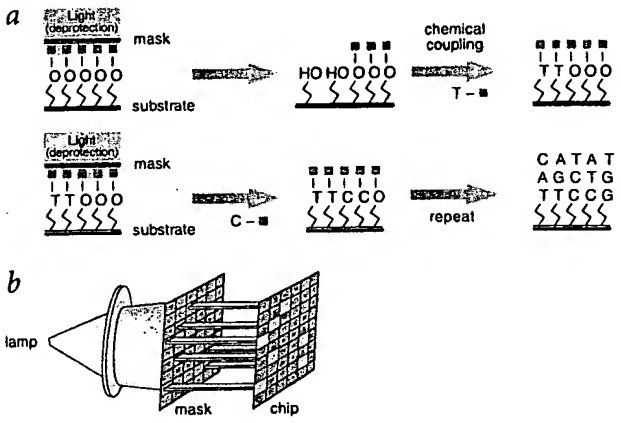


Fig. 1 ***a***, Light directed oligonucleotide synthesis. A solid support is derivatized with a covalent linker molecule terminated with a photolabile protecting group. Light is directed through a mask to deprotect and activate selected sites, and protected nucleotides couple to the activated sites. The process is repeated, activating different sets of sites and coupling different bases allowing arbitrary DNA probes to be constructed at each site. ***b***, Schematic representation of the lamp, mask and array.

Exploring the new world of the genome with DNA microarrays

Patrick O. Brown[1,3] & David Botstein[2]

Departments of [1]Biochemistry and [2]Genetics, and the [3]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California 94305, USA. e-mail: pbrown@cmgm.stanford.edu

Thousands of genes are being discovered for the first time by sequencing the genomes of model organisms, an exhilarating reminder that much of the natural world remains to be explored at the molecular level. DNA microarrays provide a natural vehicle for this exploration. The model organisms are the first for which comprehensive genome-wide surveys of gene expression patterns or function are possible. The results can be viewed as maps that reflect the order and logic of the genetic program, rather than the physical order of genes on chromosomes. Exploration of the genome using DNA microarrays and other genome-scale technologies should narrow the gap in our knowledge of gene function and molecular biology between the currently-favoured model organisms and other species.

The genome project has revitalized exploration in biological research. Not long ago, it was possible for biologists to imagine that the genes that had been discovered via mutations, selections and cloning schemes represented a good approximation of the total universe of genes, and that the proteins already discovered on the basis of their abundance, location, or activity well represented the total universe of proteins. One of the great contributions of the genome project has been to show us what a small part of this world was really known to us, and how much of this world remains to be explored. In April 1996, the complete sequence of the yeast genome confronted us with the fact that yeast contain approximately 6,200 'real' genes, as judged from open reading frames, for only one quarter of which could we hazard a guess regarding function[1] (http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html). The tens of thousands of partial human cDNA sequences representing previously unseen genes have had a similar humbling effect[2]. Although we may have suspected its existence, the actual discovery of this genetic *terra incognita* has jolted biology much as the discovery of America jolted Europe 500 years ago—showing us how much of the world is beyond the frontier—mysterious, tantalizing and unexplored.

Exploring the genome and the natural world with DNA microarrays

Exploration means looking around, observing, describing and mapping undiscovered territory, not testing theories or models. The goal is to discover things we neither knew or expected, and to see relationships and connections among the elements, whether previously suspected or not. It follows that this process is not driven by hypothesis and should be as model-independent as possible (see page 54 of this issue (ref. 3)). We should use the unprecedented experimental opportunities that the genome sequences provide to take a fresh, comprehensive and open-minded look at every question in biology. If we succeed, we can expect that many of the new models that emerge will defy conventional wisdom.

Exploring and surveying are best done systematically. The genome, representing the complete blueprint of the organism, is the natural bounded system in which to conduct this exploration. The completion of the genomic sequences of 'model organisms' (currently the eukaryotes *Saccharomyces cerevisiae* and *Caenorhabditis elegans*, as well as dozens of bacterial species) provides us with such complete blueprints (http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html). These genome sequences have not only made a new era of exploration imperative, but, providentially, they have also made it possible.

DNA microarrays provide a simple and natural vehicle for exploring the genome in a way that is both systematic and comprehensive[4–10]. The power and universality of DNA microarrays as experimental tools derives from the exquisite specificity and affinity of complementary base-pairing. We are provided thereby with an instant experimental handle on DNA or RNA unlike any we possess for any other biological molecules. A DNA copy of an individual gene provides a nearly ideal reagent for specific and quantitative detection and measurement of the sequence of the gene, even in an extremely complex mixture. For this reason, the sequence information provided by the genome project has had an instantaneous impact on experimental biology.

The method used in our labs is simple to describe (complete details and protocols are available, http://cmgm.stanford.edu/pbrown). Briefly, arrays of thousands of discrete DNA sequences (for example, all of the 6,200 known and predicted genes of *S. cerevisiae*) are printed on glass microscope slides using a robotic 'arrayer' (ref. 5; see also pages 10 (ref. 11) and 15 (ref. 12) of this issue). To compare the relative abundance of each of these gene sequences in two DNA or RNA samples (for example, the total mRNA isolated from two different cell populations; Fig. 1), the two samples are first labelled using different fluorescent dyes (say, a red dye and a green dye). They are then mixed and hybridized with the arrayed DNA spots. Use of differentially labelled mixtures avoids most of the complications of hybridization kinetics; we always measure the ratio. After hybridization, fluorescence measurements are made with a microscope that illuminates each DNA spot and measures fluorescence for each dye separately; these measurements are used to determine the ratio, and in turn the relative abundance, of the sequence of each specific gene in the two mRNA or DNA samples. There are, of course, other microarray systems and methods, most notably the oligonucleotide arrays developed by Affymetrix[6,7,13], which differ in many details but share the essential simplicity of this experimental design.

1998 Feb 26 8(5) R171-4

Gene chips: **Array of hope for understanding gene regulation**

Mark Johnston

High density arrays of DNA fragments on a solid surface allow the expression of thousands of genes to be assessed in a single experiment. The development of this 'gene chip' technique heralds a new era of studies that promises to provide an integrated view of the expression of all genes of an organism.

Address: Department of Genetics Box 8232, Washington University Medical School, 4566 Scott Avenue, St. Louis, Missouri 63110, USA.

Current Biology 1998, 8:R171–R174
http://biomednet.com/elecref/09609822008R0171

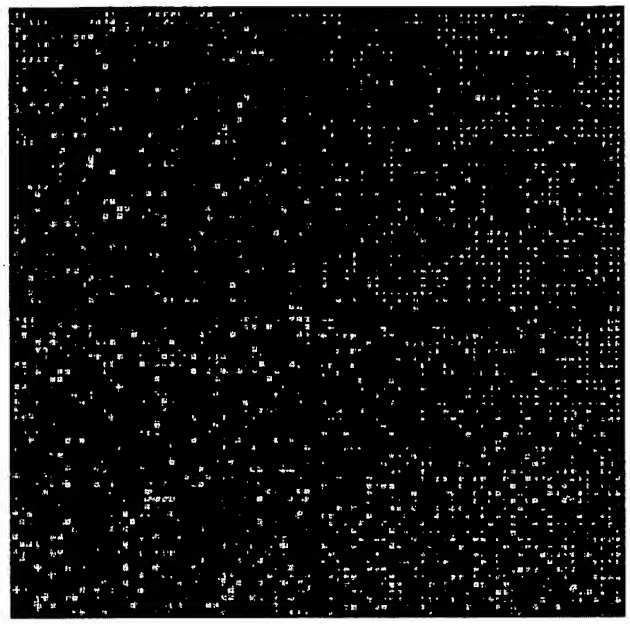
The realization almost a half century ago that genes that are co-regulated often encode proteins of related function fueled a remarkable period of discovery about mechanisms of gene regulation. The paradigms provided by this work form a cornerstone of molecular biology [1]. These paradigms were verified and extended by painstaking dissection of regulatory mechanisms, operon by operon (for eukaryotes, regulon by regulon). While 'global' regulatory mechanisms that act upon large sets of genes were recognized early on [2], their analysis has been, for the most part, limited to a small number of representative genes subject to global control. Our present knowledge of how genes are regulated thus stems from analysis of a limited number of genes. A revolutionary new technology for measuring expression of all genes of an organism in a single experiment has now been devised [3–5]. This may herald a new era of investigation of gene regulation that promises to provide a much deeper understanding of how cells coordinate expression of thousands of genes.

This advance was made possible by the development of technology that allows DNA fragments to be arrayed at high density on a solid support for use in hybridization experiments [6,7]. Thousands of DNA fragments can be arrayed on a surface no larger than a fingernail and used to probe the mRNA content of cells. Thus, whole genomes can be assessed for their pattern of gene expression, enabling us, for the first time, to view gene regulation in the context of all the complex networks of pathways that operate in cells. We can now identify all the genes of an organism that change expression under a given condition, and hope to make sense of the cell's response to that condition. Such information can provide key clues to the function of individual proteins. Moreover, the ability to acquire data of this kind is a big step toward achieving the ultimate goal of molecular biology: a complete understanding of cellular function.

Two different methods for arraying large numbers of DNA molecules in a very small space have been developed. In one, cDNA-sized fragments — usually produced by the polymerase chain reaction (PCR) — are spotted onto polylysine-coated glass slides [6]. In the other, short (~25 nucleotide) oligonucleotides are synthesized on a glass surface [7]. The arrays produced by both methods have been called 'chips', but this moniker fits the oligonucleotide arrays better, because they are made using photolithographic masks similar to those used for fabricating computer chips. Both methods pack thousands of DNA fragments into a very small area; the current oligonucleotide chips display all 6000 yeast genes on four 1.28 × 1.28 cm chips; the DNA fragment microarrays fit the same information onto a single 1.8 × 1.8 cm glass slide (see Figure 1). On the oligonucleotide chips, each gene must be represented by several (typically 20) different oligonucleotides, because of the differences in hybridization properties and reduced hybridization specificity inherent in such short probes. In addition, each oligonucleotide on the chips has a partner adjacent to it that differs at just one central base, which serves as an internal control for hybridization specificity. Each gene thus encompasses about 40 'features' — a feature being an area of the glass surface occupied by DNA molecules of one sequence — on an oligonucleotide chip, whereas it takes up only one feature on a DNA fragment microarray.

The DNA arrays are used to interrogate complex mixtures of nucleic acids, and thus are similar to the 'dot blots' that have been in use for a long time [8,9]. They differ from dot blots in the nature of the labelled species that serves as the probe — in dot blots, the complex mixture of mRNAs is fixed to the solid surface and probed with a single labelled DNA fragment; in the DNA microarrays, individual unlabelled DNA fragments are fixed on the solid support and probed with a complex mixture of labelled cDNAs or mRNAs. The major advance of the arrays over the older technology is a significant increase in sensitivity, primarily as a result of two factors. Because the labeled probe is usually the limiting component in nucleic acid hybridization, probably the more important factor is the small area occupied by the arrays, which significantly reduces the volume of the hybridization solution — from milliliters to microliters — and thereby greatly increases the concentration of the probe. Because of the small area they occupy, sophisticated lasers and sensitive detection systems are required to measure the hybridization signals. The second factor is that the glass surface of an array generates a smaller background hybridization signal than the porous membranes used for dot blots. Both kinds of

Figure 1



A DNA fragment array of all ~ 6000 yeast genes probed with labeled cDNA made from galactose- and glucose-grown cells. Each spot (element) on the array contains a cDNA-sized DNA fragment representing one yeast coding sequence. mRNA from galactose-grown cells was converted to red-labeled cDNA (using dUTP labeled with the fluorescent dye Cy3); mRNA from glucose-grown cells was converted to green-labeled cDNA (with the dye Cy5). These two preparations of labeled cDNA were mixed and used to probe the array. Red spots bind only galactose-grown cDNA, and thus represent genes expressed only in galactose-grown cells; green spots bind only cDNA from glucose-grown cells, and therefore represent genes expressed only in glucose-grown cells. Spots containing genes expressed under both conditions hybridize to both cDNAs, and thus appear yellow. The intensity of the color of each spot (from red to green) reveals the relative expression level of genes under the two conditions. (Figure courtesy of Joe DeRisi, Vishy Iyer, and Pat Brown; for more of these images, see [17].)

microarray thus permit very sensitive detection of gene expression: currently, an mRNA present at a level less than one molecule in 100,000 can be detected, equivalent to a transcript present at only one copy per 20 yeast cells!

The DNA fragment microarrays can be produced by anybody with the ability and modest means required to assemble the equipment to print the arrays [10]. Production of the DNA fragments to be arrayed does, however, require a large number of oligonucleotides for the PCR, which can be prohibitively expensive, and generation of the PCR products is labor intensive. (For yeast, much of this work has already been done [11].) A limitation of the oligonucleotide chips is that knowledge of the DNA sequences to be studied is necessary to produce them, whereas random cDNA clones can be used in the DNA fragment microarrays. Also, dependence on commercial sources for the oligonucleotide chips may present limitations of availability and affordability. Both methods require fairly sophisticated microscopy and software for detecting, measuring and identifying hybridization signals from the arrays. This technology currently seems out of the reach of the average lab, but commercial services are sprouting to provide the microarrays and equipment necessary to make this technology widely accessible. In the meantime, the whole genome dot blots that have recently become available, at least for yeast, may fulfil the needs of most labs that want to perform these kinds of experiments [12].

The utility of the two kinds of microarray for measuring expression of a large number of genes was established previously [6,13–15], but was spectacularly demonstrated recently by two groups who used them to measure expression of all 6000 genes of the bakers' yeast, *Saccharomyces cerevisiae*, grown under a few different conditions [3,4]. Wodicka *et al.* [4] compared gene expression in yeast cells grown on rich and minimal media. They isolated polyA$^+$ RNA from cells grown under the two conditions, converted it into cDNA flanked by a promoter for T7 RNA polymerase, and copied it into antisense, biotin-labeled RNA by transcription *in vitro*. This final step amplifies the mRNA probe, apparently without introducing significant bias. Labeled RNA made in this way from cells grown under the two conditions was used to probe the oligonucleotide chip, and the bound RNA was detected and quantified using streptavidin conjugated to a fluorescent dye, yielding highly reproducible results. More than 87% of yeast mRNAs were detected, with a dynamic range of about three orders of magnitude.

Similar results were obtained by DeRisi *et al.* [3], who used the DNA fragment microarrays to measure gene expression in yeast cells as they run out of glucose. They isolated polyA$^+$ RNA from a culture of cells at several different times after inoculation into glucose media, fluorescently labeled it by reverse transcription, and used the labelled product to probe DNA fragment microarrays. The two types of array seem roughly comparable in their sensitivity, range and reproducibility. The oligonucleotide chips may be better at measuring relative expression differences, because they easily revealed more than 50-fold differences in expression, whereas the maximum expression difference measured with the DNA fragment microarrays was 20-fold (although it is difficult to compare the results of the two experiments, as they employed very different growth conditions).

The results presented by DeRisi *et al.* [3] and Wodicka *et al.* [4] mostly serve to validate the experimental approach, but in a very satisfying way, as many of the changes in gene expression that were observed were expected. DeRisi *et al.* [3], for example, rediscovered the fact that, when yeast cells run out of glucose, the expression of

genes for oxidative metabolism and gluconeogenesis increases, and the expression of genes for fermentation and protein synthesis decreases. That these results conform almost perfectly to what is known about regulation of these well-studied genes lends great confidence to the technique.

Similarly, many of the genes expected to have higher levels of expression in cells grown on minimal media than on rich media — such as those involved in nitrogen acquisition or amino acid synthesis — were identified with the oligonucleotide chips, as were many genes that have the converse expression pattern, such as those involved in amino-acid transport. The technique is not perfect, however, as the DNA fragment microarrays missed several genes whose expression is known to be regulated by glucose — for example, *HXT1*, which is induced about 300-fold by glucose [16], and *GAL4*, which is about 75% repressed by glucose [17]. (These omissions could be easily uncovered, because all the results are publicly available in a terrific, searchable database [18].) Nevertheless, the microarrays work better than most of us imagined they would, and provide a wonderful tool that greatly expands our horizons.

What have these experiments taught us about cellular function? They revealed that almost 90% of yeast genes are expressed, most at very low levels (69% with one or fewer mRNAs per cell) [4], but this has long been known from the classic work of Hereford and Rosbash [19]. Similarly, many of the genes DeRisi *et al.* [3] found to be regulated by glucose have long been known to be subject to such regulation. A substantial number of genes, however, were found for the first time to be regulated in these two studies, and nothing is known about a significant proportion of these. The regulatory patterns of these proteins thus provide a first clue to their function. These results also allow genes to be grouped by their expression pattern, as was done insightfully by DeRisi *et al.* [3]. The function of at least some of the genes in a group is usually known, allowing inferences to be made about the possible function of the other genes in the same group. Clearly, this technology will speed the pace of discovery of protein and cellular function.

One of the most promising applications of DNA microarrays is the identification of all the genes whose expression changes when a gene is inactivated. This is a boon for those interested in transcription factors, as this information should help reveal their role in cellular physiology, and might even speak to their mechanism of action. DeRisi *et al.* [3] identified all yeast genes whose expression changes when the Tup1 transcription factor is inactivated by mutation. The expression of many genes increased significantly as a result of deletion of *TUP1*, which would probably lead one to conclude correctly that Tup1 is a general repressor. Interestingly, expression of a few genes decreased significantly in a *tup1* mutant, suggesting that Tup1 may also activate transcription in certain cases. In a separate set of experiments, genes whose expression changes when the Yap1 transcription factor is overexpressed were identified. This revealed a set of genes whose expression increased, indicating that Yap1 is a transcriptional activator. Again, expression of a few genes decreased significantly upon Yap1 overexpression, suggesting that Yap1 may also be a repressor.

A major problem with interpretation of these results is the difficulty in ascribing them to direct action of the transcription factor that is inactivated. In fact, it seems a good bet that indirect effects account for the unexpected responses to Tup1 absence and Yap1 overexpression. Nevertheless, the wealth of data provided by the microarrays allows the formulation of hypotheses that can be tested with other, more conventional experiments. The practical uses of this technology to identify candidate compounds for drug development are obvious. Furthermore, the microarrays are sure soon to be in wide clinical use, where they will undoubtedly aid in disease diagnosis and treatment.

Now that the DNA microarrays are clearly working well for the analysis of gene expression, the major challenge is to handle and interpret the massive amounts of data that will quickly accrue. Just from the two reports of DeRisi *et al.* [3] and Wodicka *et al.* [4], there is a rich vein of information waiting to be mined that is sure to grow as this technology becomes widely available. But the problem we are faced with is a pleasant one: we are not limited by the amount of data we can collect, but by our ability to interpret it. If we are able to do so successfully, great insight into cellular function is promised. It is unlikely to change our paradigms, but it will take us one large step closer to the goal of a complete understanding of how cells work.

Acknowledgements

I thank Stan Fields for insightful comments on the manuscript, and Joe DeRisi, Vishy Iyer and Pat Brown for generously providing the image of the yeast genome DNA fragment array.

References

1. Jacob F, Monod J: Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 1961, 3:318-356.
2. Magasanik B: Catabolite repression. *Cold Spring Harbor Symp Quant Biol* 1961, 26:249-256.
3. DeRisi JL, Iyer VR, Brown PO: Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997, 278:680-686.
4. Wodicka L, Dong H, Mittmann M, Ho M-H, Lockhart DJ: Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nat Genet* 1997, 15:1359-1367.
5. Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile C, Hwang SY, Brown PO, Davis RW: Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci USA* 1997, 94:13057-13062.
6. Schena M, Shalon D, Davis RW, Brown PO: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995, 270:467-470.

1/24/97

7. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D: **Light-directed, spatially addressable parallel chemical synthesis.** *Science* 1991, **251**:767-773.
8. Kafotos FC, Jones CW, Efstratiadis A: **Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure.** *Nucleic Acids Res* 1979, **7**:1541-1552.
9. Chuang SE, Daniels DL, Blattner FR: **Global regulation of gene expression in *Escherichia coli*.** *J Bacteriol* 1993, **175**:2026-2036.
10. **The microarray home page.** http://cmgm.stanford.edu/pbrown/array.html
11. Hudson JR Jr, Dawson EP, Rushing KL, Jackson CH, Lockshon D, Conover D, Lanciault C, Harris JR, Simmons SJ, Rothstein R, Fields S: **The complete set of predicted genes from Saccharomyces cerevisiae in a readily usable form.** *Genome Res* 1997, **7**:1169-1173.
12. **Gene filters.** http://www.resgen.com/online_catalog/cat_view_group.html?group=151009&item_id_no=-1
13. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL: **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nat Biotechnol* 1996, **14**:1675-1680.
14. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW: **Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.** *Proc Natl Acad Sci USA* 1996, **93**:10614-10619.
15. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM: **Use of a cDNA microarray to analyse gene expression patterns in human cancer.** *Nat Genet* 1996, **14**:457-460.
16. Ozcan S, Johnston M: **Three different regulatory mechanisms enable yeast hexose transporter (*HXT*) genes to be induced by different levels of glucose.** *Mol Cell Biol* 1995, **15**:1564-1572.
17. Griggs DW, Johnston M: **Regulated expression of the GAL4 activator gene in yeast provides a sensitive genetic switch for glucose repression.** *Proc Natl Acad Sci USA* 1991, **88**:8597-8601.
18. **Exploring the metabolic and genetic control of gene expression on a genomic scale.** http://cmgm.stanford.edu/pbrown/explore/index.html
19. Hereford LM, Rosbash M: **Number and distribution of polyadenylated RNA sequences in yeast.** *Cell* 1977, **10**:453-462.

Fischer-Vize, *Science* **270**, 1828 (1995).
35. T. C. James and S. C. Elgin, *Mol. Cell Biol.* **6**, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 263 (1991); B. Tschiersch *et al.*, *EMBO J.* **13**, 3822 (1994); M. T. Madireddi *et al.*, *Cell* **87**, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 7137 (1996).
36. P. M. Palosaari *et al.*, *J. Biol. Chem.* **266**, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E. Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* **58**, 4068 (1992); V. Sharma, K. Suvarna, R. Meganathan, M. E. Hudspeth, *J. Bacteriol.* **174**, 5057 (1992); M. Kanazawa *et al.*, *Enzyme Protein* **47**, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* **178**, 3015 (1996).
37. M. Ho *et al.*, *Cell* **77**, 869 (1994).
38. W. Hendriks *et al.*, *J. Cell Biochem.* **59**, 418 (1995).
39. We thank H. Skaletsky and F. Lewitter for help with sequence analysis; Lawrence Livermore National Laboratory for the flow-sorted Y cosmid library; and P. Bain, A. Bortvin, A. de la Chapelle, G. Fink, K. Jegalian, T. Kawaguchi, E. Lander, H. Lodish, P. Matsudaira, D. Menke, U. RajBhandary, R. Reijo, S. Rozen, A. Schwartz, C. Sun, and C. Tilford for comments on the manuscript. Supported by NIH.

28 April 1997; accepted 9 September 1997

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (*1*, *2*), provide a practical and economical tool for studying gene expression on a very large scale (*3*–*6*).

Saccharomyces cerevisiae is an especially favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, *cis* regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (*7*). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (*8*). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by using a simple robotic printing device (*9*). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (*10*). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (*11*) and then hybridized to the microarrays (*12*). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (*13*).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (*14*). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305–5428, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

to any gene whose function is known (*15*). The responses of these previously uncharacterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (*ALD2*) and acetyl–coenzyme A(CoA) synthase (*ACS1*), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes *PCK1*, encoding phosphoenolpyruvate carboxykinase, and *FBP1*, encoding fructose 1,6-biphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c–related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (*13*). A noteworthy and illuminating exception was that the genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitchondrial biogenesis (*13*). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (*16–20*). A search in the promoter regions of the remaining two genes, *ACR1* and *IDP2*, revealed that *ACR1*, a gene essential for *ACS1* activity, also possessed a consensus CSRE motif, but interestingly, *IDP2* did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception

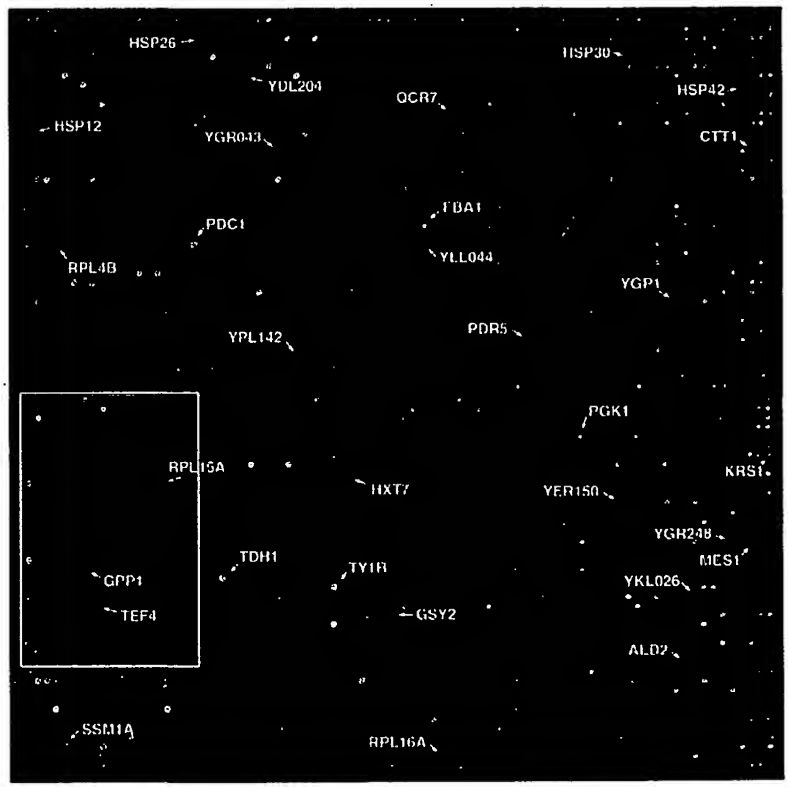


Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (*9*). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (*49*). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of $<5 \times 10^6$ cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of $\sim 2 \times 10^8$ cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP–labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP–labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of *HSP42*, have previously been shown to be controlled at least in part by these elements (*21–24*). Inspection of the sequences upstream of *HSP42* and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including *HSP30*, *ALD2*, *OM45*, and 10 uncharacterized ORFs (*25*)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex *HAP2,3,4* has been shown to be responsible for induction of several genes important for respiration (*26–28*). This complex binds a degenerate consensus sequence known as the CCAAT box (*26*). Computer analysis, using the consensus sequence TNRYTGGB (*29*), has suggested that a large number of genes involved in respiration may be specific targets of *HAP2,3,4* (*30*). Indeed, a putative *HAP2,3,4* binding site could be found in the sequences upstream of each of the seven cytochrome c–related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome c–related genes that were induced, *HAP2,3,4* binding sites were present in all but one. Significantly, we found that transcription of *HAP4* itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS_{rpg}) that is recognized by the Rap1 DNA-binding protein (*31*, *32*). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (*33*). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (*34*). Indeed, we observed that the abundance of *RAP1* mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, *HAP4* and *SIP4*, were induced by a factor of more than threefold at the diauxic shift. *SIP4* encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the "master regulator" of glucose repression (*35*). The eightfold induction of *SIP4* upon depletion of glucose strongly suggests a role in the induction of downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (*36*). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expression ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (*37*, *38*). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

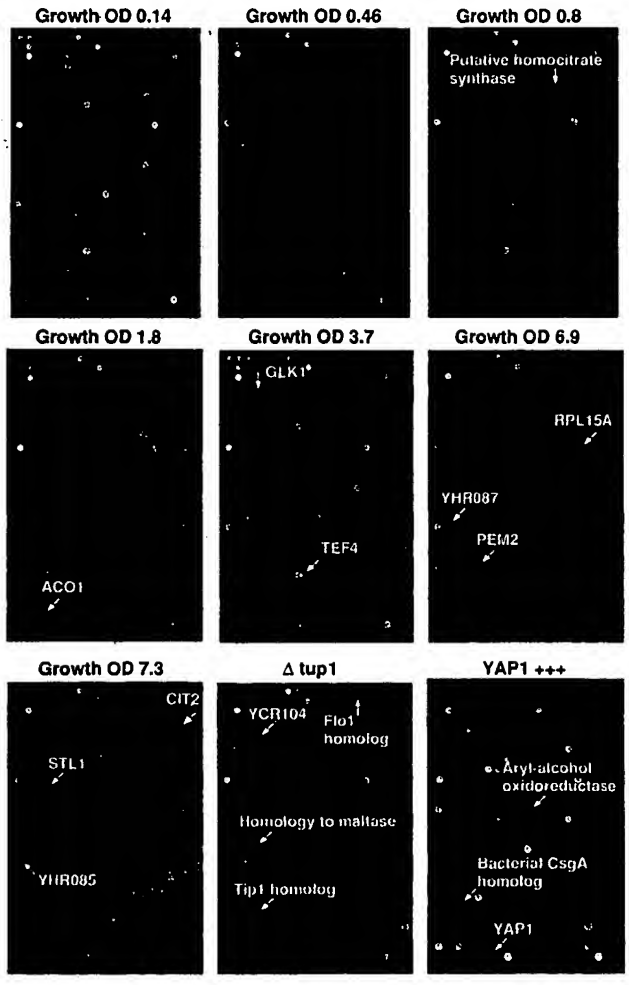


Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the *tup1*Δ mutation and *YAP1* overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (*13*). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.

by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion of the *TUP1* gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (39). Tup1 has also been implicated in repression of oxygen-regulated, mating-type–specific, and DNA-damage–inducible genes (40).

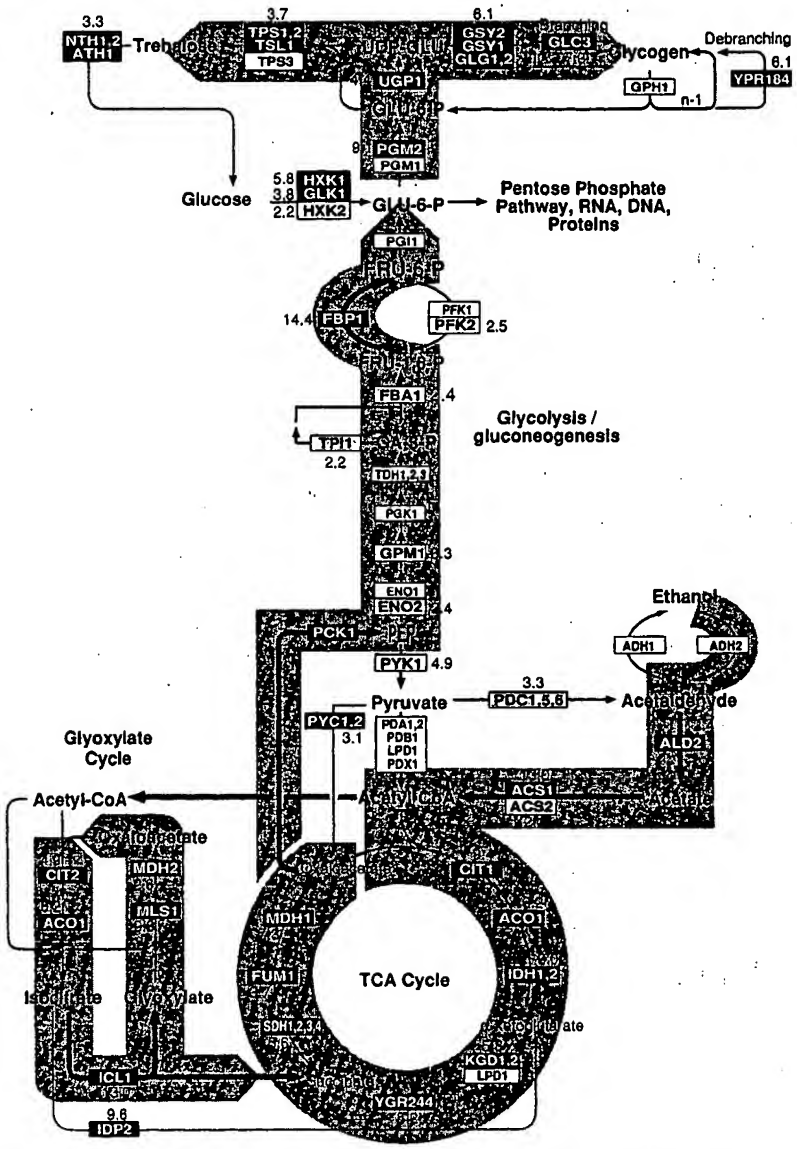


Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs YGR244 and YPR184 show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

Wild-type yeast cells and cells bearing a deletion of the *TUP1* gene (*tup1*Δ) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (*11*). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1*Δ strain, and thus presumably repressed by Tup1 (*41*). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1*Δ mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (*13*)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included α-glucosidases, the mating-type–specific genes *MFA1* and *MFA2*, and the DNA damage–inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1*Δ strain, providing a positive control in the experiment (*42*).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (*15*) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as Tip1 and Tir1/Srp1 which are induced by cold shock and other stresses (*43*), and similar, serine-poor proteins, the seripauperins (*44*). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1*Δ

strain, and 18 of these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators of the cell division cycle were induced more than twofold by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a MAT α strain in which *MFA1* and *MFA2*, the genes encoding the a-factor mating pheromone precursor, are normally repressed. In the isogenic *tup1Δ* strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a MATA strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the b-zip class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, o-phenanthroline, heavy metals, and osmotic stress (*45*). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GAL1-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

Of the 17 genes whose mRNA levels increased by more than threefold when *YAP1* was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (*46*, *47*). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing Yap1. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for Yap1-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (*48*). About two-thirds of the genes that were induced by more than threefold upon Yap1 overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by Yap1. The absence of canonical Yap1-bind-

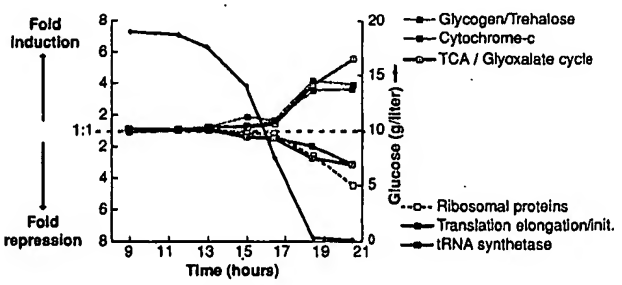


Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.

Table 1. Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (*50*). Positions of the canonical Yap1 binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of Yap1 site from ATG	Gene	Description	Fold-increase
YNL331C			Putative aryl-alcohol reductase	12.9
YKL071W	162–222 (5 sites)		Similarity to bacterial csgA protein	10.4
YML007W		YAP1	Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W			Putative aryl-alcohol reductase	6.5
YML116W	409	ATR1	Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	OYE3	NAPDH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	OYE2	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to *A. thaliana* zeta-crystallin homolog	3.7
YOL126C		MDH2	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and networks. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

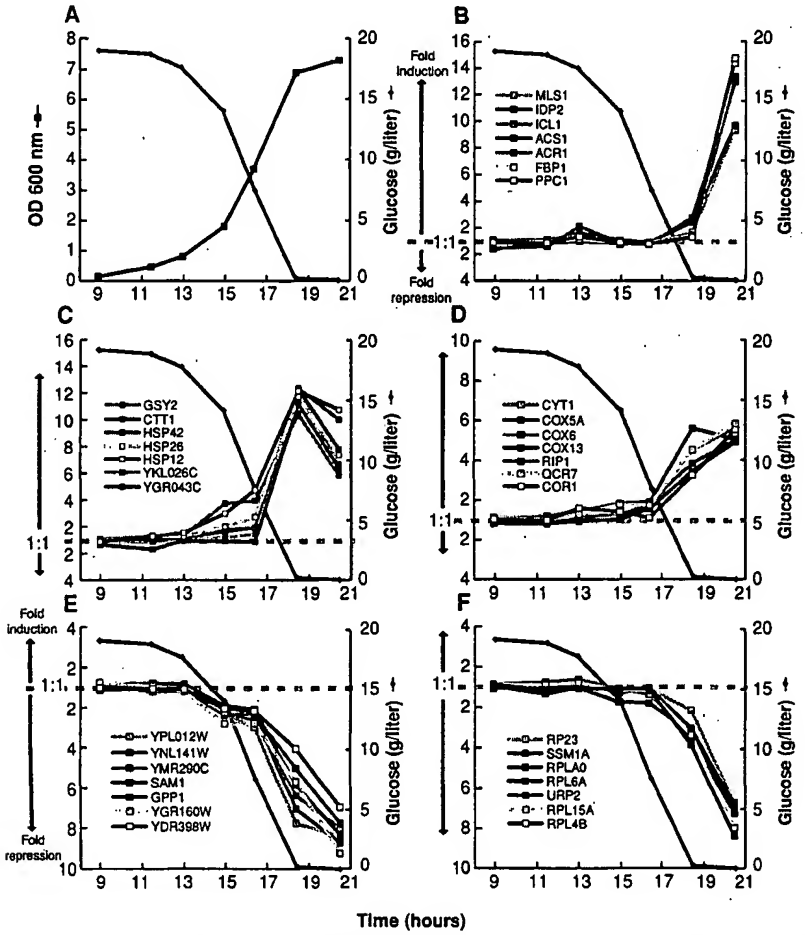


Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (**A**) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (**B**) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (**C**) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contain STRE motif repeats in their upstream promoter regions. (**D**) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (**E**) *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (**F**) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* **6**, 639 (1996).
3. D. Lashkari, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
4. J. DeRisi *et al.*, *Nature Genet.* **14**, 457 (1996).
5. D. J. Lockhart *et al.*, *Nature Biotechnol.* **14**, 1675 (1996).
6. M. Chee *et al.*, *Science* **274**, 610 (1996).
7. M. Johnston and M. Carlson, in *The Molecular Biology of the Yeast* Saccharomyces: *Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100-μl PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3× standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-

tion, the bound DNA was denatured by a 2-min incubation in distilled water at ~95°C. The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.

10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251) Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at −80°C.
11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 μg of polyadenylated [poly(A)+] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 μM for dATP, dCTP, and dGTP and 200 μM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100 μM. The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with of 470 μl of 10 mM tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to ~5 μl, using Centricon-30 microconcentrators (Amicon).
12. Purified, labeled cDNA was resuspended in 11 μl of 3.5× SSC containing 10 μg poly(dA) and 0.3 μl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ~8 to 12 hours in a water bath at 62°C. Before scanning, slides were washed in 2× SSC, 0.2% SDS for 5 min, and then 0.05× SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.
13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html
14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.
15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: *Saccharomyces* Genome Database (genome-www.stanford.edu), Yeast Protein Database (quest7.proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/mips/yeast/index.htmlx).
16. A. Scholer and H. J. Schuller, *Mol. Cell. Biol.* **14**, 3613 (1994).
17. S. Kratzer and H. J. Schuller, *Gene* **161**, 75 (1995).
18. R. J. Haselbeck and H. L. McAlister, *J. Biol. Chem.* **268**, 12116 (1993).
19. M. Fernandez, E. Fernandez, R. Rodicio, *Mol. Gen. Genet.* **242**, 727 (1994).
20. A. Hartig *et al.*, *Nucleic Acids Res.* **20**, 5677 (1992).
21. P. M. Martinez *et al.*, *EMBO J.* **15**, 2227 (1996).
22. J. C. Varela, U. M. Praekelt, P. A. Meacock, R. J. Planta, W. H. Mager, *Mol. Cell. Biol.* **15**, 6232 (1995).
23. H. Ruis and C. Schuller, *Bioessays* **17**, 959 (1995).
24. J. L. Parrou, M. A. Teste, J. Francois, *Microbiology* **143**, 1891 (1997).
25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.
26. S. L. Forsburg and L. Guarente, *Genes Dev.* **3**, 1166 (1989).
27. J. T. Olesen and L. Guarente, *ibid.* **4**, 1714 (1990).
28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, *Mol. Microbiol.* **13**, 119 (1994).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B–C, G, or T; N–G, A, T, or C; R–A or G; and Y–C or T.
30. C. Fondrat and A. Kalogeropoulos, *Comput. Appl. Biosci.* **12**, 363 (1996).
31. D. Shore, *Trends Genet.* **10**, 408 (1994).
32. R. J. Planta and H. A. Raue, *Ibid.* **4**, 64 (1988).
33. The degenerate consensus sequence VYCYRNNC-MNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (*30*), is WACAYCCRTACATYW, with up to three differences allowed.
34. S. F. Neuman, S. Bhattacharya, J. R. Broach, *Mol. Cell. Biol.* **15**, 3187 (1995).
35. P. Lesage, X. Yang, M. Carlson, *Ibid.* **16**, 1921 (1996).
36. For example, we observed large inductions of the genes coding for *PCK1*, *FBP1* [Z. Yin *et al.*, *Mol. Microbiol.* **20**, 751 (1996)], the central glyoxylate cycle gene *ICL1* [A. Scholer and H. J. Schuller, *Curr. Genet.* **23**, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, *ACS1* [M. A. van den Berg *et al.*, *J. Biol. Chem.* **271**, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes *PYK1* and *PFK2* [P. A. Moore *et al.*, *Mol. Cell. Biol.* **11**, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase T *CTT1* [P. H. Bissinger *et al.*, *ibid.* **9**, 1309 (1989)] and several genes encoding small heat-shock proteins, such as *HSP12*, *HSP26*, and *HSP42* [I. Farkas *et al.*, *J. Biol. Chem.* **266**, 15602 (1991); U. M. Praekelt and P. A. Meacock, *Mol. Gen. Genet.* **223**, 97 (1990); D. Wotton *et al.*, *J. Biol. Chem.* **271**, 2717 (1996)].
37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, *FBP1* and *PCK1*) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (*46*).
38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, *ADH1* and *ADH2*, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as *HXK1/HXK2* (77% identical) [P. Herrero *et al.*, *Yeast* **11**, 137 (1995)], *MLS1/DAL7* (73% identical) (*20*), and *PGM1/PGM2* (72% identical) [D. Oh, J. E. Hopper, *Mol. Cell. Biol.* **10**, 1415 (1990)], In accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.
39. F. E. Willams, U. Varanasi, R. J. Trumbly, *Mol. Cell. Biol.* **11**, 3307 (1991).
40. D. Tzamarias and K. Struhl, *Nature* **369**, 758 (1994).
41. Differences in mRNA levels between the tup1Δ and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concordance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the tup1Δ strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.
42. The tup1Δ mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of *TUP1* and an Eco R I site 124 base pairs before the stop codon of the *TUP1* gene.
43. L. R. Kowalski, K. Kondo, M. Inouye, *Mol. Microbiol.* **15**, 341 (1995).
44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, *Gene* **148**, 149 (1994).
45. D. Hirata, K. Yano, T. Miyakawa, *Mol. Gen. Genet.* **242**, 250 (1994).
46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, *Appl. Environ. Microbiol.* **60**, 1783 (1994).
47. A. Muheim *et al.*, *Eur. J. Biochem.* **195**, 369 (1991).
48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, *J. Biol. Chem.* **269**, 32592 (1994).
49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm.stanford.edu/pbrown. Images were scanned at a resolution of 20 μm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.
50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold)
51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the Tup1 deletion strain; L. Fernandes for helpful advice on Yap1; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997